# Portrait of a Pathogen:

A characterisation of *Mycobacterium tuberculosis* and its host using multi-dimensional proteomics.




James Luke Gallant

UNIVERSITET STELLENBOSCH

# Portrait of a Pathogen:
## A characterisation of Mycobacterium tuberculosis and its host using multi-dimensional proteomics

Deur

**James Luke Gallant**

Bsc Molekulêre Biologie en Biotegnologie, Stellenbosch Universiteit, 2012

BscHons Molekulêre Biologie, Stellenbosch Universiteit, 2013

Msc Molekulêre Biologie, Stellenbosch Universiteit, 2016

*Proefskrif ingelewer vir die graad Doktor in Molekulêre Biologie in die Fakulteit Geneeskunde en Gesondheidswetenskappe aan die Universiteit Stellenbosch. Hierdie proefskrif is ook ingedien by die Vrije Universiteit Amsterdam in terme van 'n dubbelegraad-oorenkoms.*

Supervisor: **Prof S.L. Sampson, Prof. Dr. W. Bitter**

Co-supervisor: **T.D.J. Heunis**

December 2021

**VERKLARING**

Deur hierdie proefskrif elektronies in te lewer, verklaar ek dat die geheel van die werk hierin vervat, my eie, oorspronklike werk is, dat ek die alleenouteur daarvan is (behalwe in die mate uitdruklik anders aangedui), dat die reporduksie daarvan deur die Universiteit van Stellenbosch nie derdepartyregte sal skend nie en dat ek dit nie vantevore in die geheel of gedeeltelik ter verkrygin van enige kwalifikasie aangebied het nie. Hierdie tesis is ook ingedien by die Vrije Universiteit Amsterdam in terme van ’n dubbelgraad-ooreenkoms. 01/12/2021

Hierdie tesis sluit 5 oorspronklike artiekels gepubliseer in ewekniebeoordeelde vak-wetenskaplike tydskrifte en 3 ongepubliseerde werke in. Die ontwikkeling en skryf van die was hoofsaaklik my eie werk en vir elkeen van die artikels waar dit nie die geval is nie is ‘ verklaring in die proefskrif ingesluit wat die aard en omvang van mede-outeurs se bydrae aandui.

VRIJE UNIVERSITEIT

# Portrait of a Pathogen:

A characterisation of *Mycobacterium tuberculosis* and its host using multi-dimensional proteomics.

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. C.M. van Praag,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op woensdag 13 oktober 2021 om 11.45 uur
in een bijeenkomst van de universiteit
De Boelelaan 1105

Door

James Luke Gallant

geboren te Kaapstad, Zuid Afrika

| | |
|---|---|
| promotoren: | prof.dr. W. Bitter |
| | prof.dr. S.L. Sampson |
| copromotor: | dr. T. de Jager Heunis |

# TABLE OF CONTENTS

# 1

# General introduction

This chapter presents an overview of tuberculosis as a global and ancient pandemic. Specifically how the disease, and it's aetiologic agent *Mycobacterium tuberculosis,* has influenced research and development in both medicine and infectious disease as a study topic, will be discussed. A summary of the core technologies or topics that underpin the research chapters of this thesis will be introduced here as well. These include the use of next generation sequencing technologies, the unique evolution of *M. tuberculosis*, the composition of the bacterial cell wall and associated secretion systems as well as the host response to *M. tuberculosis*.

## A BRIEF HISTORY OF TUBERCULOSIS

*M. tuberculosis* is the causative agent of tuberculosis, a major current global health problem. Tuberculosis is also an ancient disease with palaeopathological evidence dating back to the Neolithic period (~8000 BCE) (1). It is therefore safe to say that tuberculosis has plagued humans for centuries. This pathogen has been referred to by various names such as: King's evil, consumption, phthisis, captain of these men of death, the white plague and played an important role in shaping modern society (2–4).

Ancient civilisations had to contend with infection by *M. tuberculosis* and naturally attempted to study the disease with the knowledge and world view of the time. While notable advancements were made in the 19$^{th}$ century, ancient scientists were able to rationalise remarkable characteristics of tuberculosis for the time period. The first appearance of *M. tuberculosis* in historical record dates back to the ancient Egyptian civilisation, where Egyptian mummies have been identified with Pott's lesions (an extrapulmonary form of tuberculosis that affected the vertebrae, resulting in malformations of the spine) in approximately 2400 BC (5). While the Egyptians seemed to be aware of a fatal disease in their midst, there was little to no notable documentation of a tuberculosis-like disease in the papyrus that has survived to the modern age (6). Scholars first wrote about a disease resembling tuberculosis approximately 3000 years ago as evident by text discovered in modern day India and China (7). The disease is also mentioned in two books of the old testament, namely Leviticus 26:16 and Deuteronomy 28:22. The ancient Greeks had a remarkable grasp on tuberculosis as an infectious and fatal disease as evident by their name for tuberculosis, namely phthisis (wasting away) also referred to as consumption or "tering". This name for tuberculosis would be the dominant name until the 19$^{th}$ century. Below is the account of Hippocrates as he described tuberculosis in his book *Of the epidemics* as translated by Francis Adams (8).

*"The greatest and most dangerous disease, and the one that proved fatal to the greatest number, was consumption. With many persons it commenced during the winter, and of these some were confined to bed, and others bore up on foot; the most of those died early in spring who were confined to bed; of the others, the cough left not a single person, but it became milder through the summer; during the autumn, all these were confined to bed, and many of them died, but in the greater number of cases the disease was long protracted. Most of these were suddenly attacked with these diseases, having frequent rigors, often continual and acute fevers; unseasonable, copious, and cold sweats throughout; great coldness, from which they had great difficulty in being restored to heat; the bowels variously constipated, and again immediately in a loose state, but towards the termination in all cases with violent looseness of the bowels; a determination downwards of all matters collected about the lungs; urine excessive, and not good, troublesome melting. The coughs throughout were frequent, and copious, digested, and liquid, but not brought up with much pain; and even when they had some slight pain, in all cases the purging of the matters about the lungs went on mildly. The fauces were not very irritable, nor were they troubled with any saltish humors; but there were viscid, white, liquid, frothy, and copious defluxions from the head. But by far the greatest mischief attending these and the other complaints, was the aversion to food, as has been described. For neither been described. For neither had they any relish for drink along with their food, but continued without thirst. There was heaviness of the body, disposition to coma, in most cases swelling, which ended in dropsy; they had rigors, and were delirious towards death." - Hippocrates*

It becomes apparent that Hippocrates had a firm grasp on the symptoms of tuberculosis and could recognise the afflicted by their symptoms, which were masterfully described. However, the method by which consumption was acquired was lost on Hippocrates, as he guessed a hereditary component. This was however not the only contribution to tuberculosis research made by the ancient Greeks. Both Isocrates and Aristotle were aware of the ability of tuberculosis to spread from one individual to another and touched upon the airborne infectious nature of tuberculosis.

"*Why are those taken by phthisis, who are brought in contact with the sufferer, and not taken by such diseases as dropsy, fever and apoplexy, however close the contact with sufferers from this disease may be? Phthisis spoils the air and makes it dangerous, thus others become infected." -Isocrates*

*"Why when one comes near consumptives, does one contract their disease? The reason is that the breath is bad and heavy, one breathes this pernicious air and takes in the disease because there is in the air something disease producing" -Aristotle*

The Roman empire added their contributions to tuberculosis research as well. The Romans were able to diagnose tuberculosis and even offered a cure to this ailment (9). Furthermore the Byzantine Roman empire were accurately able to describe pulmonary and glandular tuberculosis (10). Probably the greatest Roman impact on tuberculosis was the spread of the disease across the known world at the time. DNA evidence has emerged to show the spread of tuberculosis across three continents as Rome provided roads and infrastructure to connect the Mediterranean, Europe, Africa and Asia by trade routes and the famous Roman roads (11).

While some advancements were made in England and France after the Romans, the first major contribution to the study of Tuberculosis came in 1699 from Leiden University in the Dutch Republic. Francis Sylvius, a respected physician and anatomist, provided the first clear description of tubercles and their progression to lung abscesses and cavities in his thesis, *Opera Medica* (12). He also noted that these nodules found in consumptive patients resembled that of skin ulcers caused by scrofula, of which the infectious agent is *Mycobacterium scrofulaceum.* In another thesis, *Praxeos medicae idea nova,* Sylvius noted that phthisis is scrofula of the lung and thus drew a link between two distinct yet related diseases and their causative agents (13). The notion of tuberculosis as a pulmonary disease was rediscovered by Rene Theophile Hyacinthe Laennec, inventor of the stethoscope, in the early 19$^{th}$ century some 200 years after Francis Sylvius (14). Laennec's greatest contribution came from the realisation that the tubercle formation seen in various organs originate from the same disease, namely tuberculosis and thus arguing for a renaming to phthisis. The invention of the stethoscope is a direct consequence of Leannec's aversion to placing his ear upon the chest of a young woman in order to listen to her heart (14). By use of the stethoscope, a tool born from Leannec's modesty, he was able to diagnose patients with pulmonary tuberculosis. This was the first diagnosis method for tuberculosis before death. These discoveries contributed to his thesis, D'Auscultation Mediate, which is the first major body of work describing the pathology of tuberculosis as well as the physical signs of tuberculosis in living individuals (4).

The first scientific demonstration of the infectious nature of tuberculosis arose in 1865 (15). The French military surgeon, Jean-Antoine Villemin, demonstrated the infectious nature of tuberculosis by inoculating a rabbit with liquid from a tubercle cavity which he removed from a patient who had succumbed to tuberculosis (16). He further demonstrated transmission from cattle to rabbits as well as inter-rabbit transmission, however his findings were predominantly ignored at the time. Villemin however noted that infected rabbits were largely without symptoms with autopsy revealing the presence of infection (16). During the same period as Villemin's demonstration, William Bud set

forth a hypothesis that tuberculosis is spread across communities by specific germs. In his letter to The Lancet, he lists multiple observations on tuberculosis among which it is noted that phthisis (tuberculosis) does not originate spontaneously, is a zymotic disease (implies that a causative agent is present) and is propagated from one person to another (17). The medical advancements in the 19th century on the diagnoses of tuberculosis were paralleled with major strides in microbiology. Unbeknownst to those studying the disease at the time, the late 19th century would merge the two fields and link microbes to disease through the ground-breaking work of Heinrich Robert Koch

In the 19th century, the medical community were thus aware that diseases like tuberculosis are infectious and communicable. In turn the scientific community was aware that microorganisms occur naturally in the environment and that these organisms do not spontaneously generate but are ever present. Spontaneous generation was officially disproven by Robert Koch in his work with *Bacillus anthraxis* during the development of the Koch postulates in 1877 (18). This directly resulted in the establishment of germ theory and its acceptance in the scientific community. A few years later, on 24 March 1882, Robert Koch made another contribution to the understanding of infectious disease and delivered his famous presentation, Die Aetiologie der Tuberkulose. In this presentation, Koch demonstrated that Phthisis was caused by *M. tuberculosis,* by use of a systematic method for detecting the origin of infectious disease. This came to be known as the Koch postulates for determining infectious aetiology, which is to this day the standard for establishing an aetiology for an infectious disease (19). Although it was firmly established at this time that tuberculosis is caused by *M. tuberculosis,* there was no effective treatment against this ailment. Even Koch's major invention of tuberculin proved to be ineffective. In fact, the first treatment that offered some progress was artificial pneumothorax. The move towards a tuberculosis vaccination began in 1908 when Albert Calmette and Camille Guérin cultivated tubercle bacilli, obtained from the udder of a cow, at the Pasteur Institute in the city of Lille, France. During their attempts to cultivate the bacilli they counteracted the clumpy nature of mycobacteria by adding ox bile and observed that sub-culturing decreased the virulence of the bacilli. This bacilli was sub-cultured for 11 years until 1919 when Calmette and Guérin inoculated a Guinea pig and observed that no disease progression occurred (20). This built on the early work of Edward Jenner and marked the first vaccine, called Bacille Calmette Guérin (BCG), against tuberculosis. This vaccine is still a standard vaccination given in many countries shortly after birth. For the next few decades there were little to no major discoveries as the world recovered from world war 1. However a different major discovery was at hand which would revolutionise the world of medicine and kickstart tuberculosis research once again.

On the third of September 1928 Professor Alexander Flemming found that one of his staphylococci cultures was contaminated with a fungus that inhibited the growth of the bacteria. This finding was refined later by Florey and Chain and the world was introduced to penicillin, one of the first antibiotics produced by a microorganism. Wartime efforts saw a need for penicillin and mass production and distribution was achieved in 1945. While penicillin was active against a multitude of pathogens, a major pathogen that remained viable in the presence of penicillin was tuberculosis. However, the first tuberculosis antibiotic was not far behind and early trials with a new antibiotic, streptomycin, were already successful at this time. Pioneering work by Albert Schatz, Elizabeth Bugie and Selman Waksman isolated Streptomycin from *Streptomyces griseus* in 1944. A total of 10g could be isolated, which was enough to test three guinea pigs (21–23). The initial trials were successful and a further 25 Guinea pigs were tested and cured after 49 days compared to the 24 control animals in 1945 (24). This sparked an antibiotic revolution in *M. tuberculosis* research with multiple first-line chemotherapies being developed between 1944 and 1960. For the first time in the history of tuberculosis, the pathogen was under control, or so it seemed. The patients treated with streptomycin showed a very high relapse rate and only with the arrival of combination therapy tuberculosis patients could be really cured. In many ways the problem of tuberculosis was considered solved and public interest in this disease waned. Indeed a publication in the journal *American Review of Tuberculosis and Pulmonary Diseases* by Johannes Holm clearly illustrated the sentiment of the day, that tuberculosis is curable with the antibiotics at hand (25). This sentiment lasted approximately 20 years, which in the context of tuberculosis plaguing human kind is barely a blink of an eye.

The return of Tuberculosis to the forefront of public health threats was sparked by another infectious disease which started gaining notoriety in the 1980's (26). The rise of Human Immunodeficiency Virus (HIV) and acquired immunodeficiency deficiency syndrome (AIDS) provided the catalyst for the resurgence of tuberculosis (27). Many individuals with latent tuberculosis were getting ill when their immune system was compromised by HIV/AIDS (28). Thus a large number of deaths around HIV and AIDS were not due to the virus its self but due to the absence of a functional immune system. Infection by *M. tuberculosis* and HIV/AIDS is so intertwined that research into the synergy between these diseases is ongoing to this day. This resurgence of tuberculosis infections during the 1980's to early 2000's included large scale treatments of the population with antibiotics (29). The increase in use of antibiotics resulted in the increase in antibiotic resistance to the point where many strains evolved resistance to first and second line antibiotics (30). Antibiotic resistance of *M. tuberculosis* threatens to revive this disease to the mortality rates observed before the use of chemotherapies and naturally interest and research into the disease increased dramatically.

The late 1990's and early 2000's issued revived tuberculosis research. Major advancements in molecular biology techniques were being developed, including high throughput sequencing and proteomics. The technology stemming from the human genome project was pursued by the group of Stuart Cole at the Pasteur Institute in Paris, France resulting in the publication of the first *M. tuberculosis* genome in 1998 (31). This was the beginning of the molecular investigation and appreciation of *M. tuberculosis* as an unusual and puzzling organism. The secrets of the *M. tuberculosis* genome were unlocked, paving the way for major scientific advances. Here, I will limit myself to the ones that relevant to the topics of this thesis. The first is that *M. tuberculosis* seems to have, in comparison to other bacteria, a rather unusual genome variation and evolution. Of note was the presence of a large number of related genes with high Guanine and Cytosine (GC) content which were also highly repetitive and have undergone a clonal expansion in the genome resulting in a high genomic occupancy. These genes were later described as the *pe/ppe* genes, so denoted by a conserved proline-glutamic acid or proline-proline-glutamic acid in the protein coding sequence. These genes are unique to the mycobacteria, and certain sub-families are unique to the pathogenic mycobacteria. Furthermore, no plasmids were found in the genome as well as what seems to be a complete lack of recent homologous gene transfer. With the inclusion of more genomic sequences it also became apparent that *M. tuberculosis* has a regional restriction where strains can be grouped by geographical location.

The second major point from the genomic revolution was the identification of a new protein secretion system, named the type VII secretion system in accordance with the nomenclature used in the Gram-negative bacteria (32). A handful of these secretion systems, also called ESX, are present in *M. tuberculosis*. This system was shown to be linked to the highly repetitive *pe/ppe* genes and responsible for secreting their protein products (33). Sequencing of the non-pathogenic *M. smegmatis* genome revealed that they were highly abundant only in pathogenic mycobacteria and the two major subfamilies of these proteins, i.e. the PE-PGRS and PPE-MPTR proteins, were not present in *M. smegmatis.* In addition, two members of the type VII secretion system, namely ESX-2 and ESX-5, are not present in this saprophyte either (34). Protein secretion systems allow bacteria to interact with their environment. The presence of unique systems in the pathogenic mycobacteria and this discrepancy between a major pathogen and a non-pathogenic counterpart is therefore a major topic of investigation. It was later found by Abdallah *et al* that the PE-PGRS and PPE-MPTR proteins are secreted by the ESX-5 system (33). The picture of *M. tuberculosis* type VII secretion was starting to take shape at this point where a unique method of secretion has developed to secrete a large family of unique proteins. These factors should assist the bacterium in one way or another to either cause disease or survive within its environment. While many theories

surrounding the role of these proteins have been proposed, none of them have as of yet been accepted as a definitive function (35). The main problem is simply the large number of members in the PE-PGRS and PPE-MPTR protein families. In addition, these genes and their protein products have peculiar characteristics that make their analysis complicated, including very high GC content of their genes, which makes them problematic for standard Illumina sequencing, and a very high amount of glycine residues, which is also challenging for proteomics and structural analysis. The function of PE/PPE proteins, and by extension the PE-PGRS/PPE-MPTR proteins, is therefore largely unknown. It is however hypothesised that these proteins play an important role in the modulation of the host response either by direct interaction with host processes or by interacting with the immune system. Indeed, some PE-PGRS proteins from *M. marinum* have been implicated in virulence and granuloma formation (36), while others mediate host cell apoptosis, cytokine secretion and induction of necrosis (37,38). In 2018 Ates *et al* discovered that secretion of these proteins were controlled by a single protein, PPE38. To add to the confusion, deletion of this protein occurs naturally in a number of circulating isolates, including the highly virulent Beijing strains (39). This called into question how important these proteins were to *M. tuberculosis* and forms the basis of the major investigations of this thesis.

## SHOTGUN GENOMICS AND PROTEOMICS

In this thesis, genomics; proteomics and a combination of the two, known as proteogenomics, have been implemented in multiple projects. These techniques form a central theme in the work presented here and each of these technologies will be briefly discussed in this section.

DNA forms the basis of all living organisms and acts as the storage for all structural and functional information of a cell. The central dogma of molecular biology dictates how information is processed in a living organism. The central dogma states that the primary information flow in a cell starts at DNA which is transcribed to RNA which in turn is translated to proteins. It is possible for RNA to be reverse transcribed to DNA, however once RNA has been translated to proteins the information is committed. Each step of the central dogma has associated regulatory mechanisms which increase in complexity as information moves from DNA to proteins. Genomics and proteomics has revolutionised the biological sciences by allowing unprecedented insight into cellular physiology at the molecular level. The field of genomics is concerned with studying the total genetic make up of a given organism, which provides information on evolution and traits of that organism. Proteomics is concerned with studying how the genetic ma-

terial is used to react to the environment and drive cellular function. Figure 1 serves to illustrate the importance of viewing a system within the correct context. There clearly exists a disparity between the genetic material and the proteins they encode. The genome may thus not provide the best surrogate for the function of the proteins and *vice versa*. To investigate the genome and proteome on a systems or high-throughput level, shotgun techniques are currently favoured. The shotgun sequencing technique minimises the initial complexity by breaking the DNA strands or proteins into smaller components, namely 100 bp reads or peptides. These are subsequently sequenced and reassembled by aligning the fragments to a reference. In the following paragraphs we will briefly describe shotgun genome sequencing and proteomics as well as the benefits of combining these two technologies in a cross-platform approach.

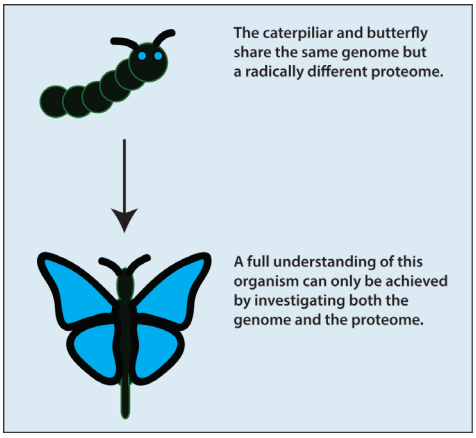



**Figure 1:** The strengths and weaknesses of genomics and proteomics.

## Genomics

Illumina sequencing is currently leading the genome sequencing market due to the high quality reads and high sequencing capacities which cannot yet be matched by long read sequencing technologies. However the contigs that are formed by short reads tend to fail in regions with duplicated genes or repetitive elements. This is especially problematic in the *pe/ppe* genes which are clonal expansions of one another and contain highly repetitive elements and are often discarded from analysis for this reason. Shotgun sequencing data can be analysed either through reference assembly, *de novo* assembly or a combination of the two can be used when called for. Reference assembly or reference mapping relies on a known reference sequence that has often been generated and completed by *de novo* assembly. Multiple paired or unpaired reads can align and the more reads that align the greater the confidence of variant calls in that region. The advantage of reference assembly is that variants that depend on a

"parent" genome can be detected, such as single nucleotide polymorphisms, deletions, insertions, etc (Figure 2, top left). The disadvantage of using a reference sequence is that alignment can fail in highly repetitive regions. In addition, reference assembly suffers from the so called reference bias. As reference sequences are often arbitrary, a single reference might not be indicative of all organisms from that species. Therefore features such as novel genes or genetic sequences cannot be detected and form part of the "dark genome". *De novo* assembly is used when a reference is not available or to detect features such as insertions which cannot be detected by reference assembly. In *de novo* assembly reads are aligned based on their overlapping base pairs to form contigs and overlapping contigs are aligned to form consensus sequences (Figure 2, top right). While *de novo* assembly can generate large consensus sequences it is rare to assemble a full genome completely with short read sequencing. The gaps between contigs are subsequently filled with low throughput techniques such as Sanger sequencing. As mentioned problematic regions can occur during reference alignment and the breakpoints are rarely clear-cut. With single nucleotide polymorphisms this does not pose a problem with enough coverage, however with structural variation unresolved breakpoints are problematic. To increase resolution, unmapped reads as well as aligned

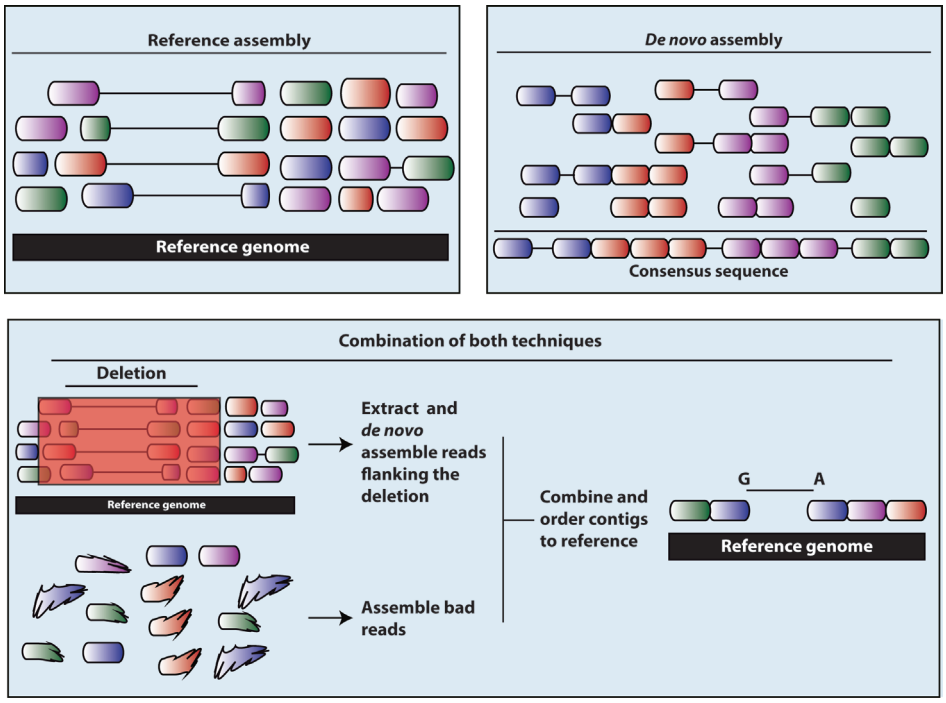



**Figure 2***:* Reference (**top left**) and *de novo* (**top right**) assembly algorithms. In this thesis a combination approach (**bottom**) was used to resolve problematic regions and gain base pair resolution on genomic break points.

reads flanking the structural variation site can be extracted, aligned and combined to create a consensus sequence (Figure 2, bottom). This reduces ambiguity and increases the resolution of breakpoints. This demonstrates the flexibility within the analysis of a genome to gain a variety of information without the need for additional sequencing.

## Proteomics

The genome is the blueprint of the cell and contains all the information to create the structural and functional processes of the cell. While *de novo* peptide assembly can be done, the technology is not yet mature and most proteomics studies still rely on database search strategies. While DNA is made up of four fundamental building blocks, proteins are much more complex molecules that can be built from 20 different amino acids. Thus the processes involved in shotgun proteomics are not universally applicable. As per example, if a protein has few or no trypsin sites, this protein will not be detected by tandem mass spectrometry. This can be mitigated by additional proteases or complementary studies, but often the problem persists and total coverage cannot be guaranteed. Due to this and a multitude of other reasons the peptide spectral matches achieved by proteomics have a much lower efficiency (~35-45%) than genome alignment scores (95%-98%). Nonetheless a standard workflow for shotgun proteomics has been consolidated as shown in figure 3.

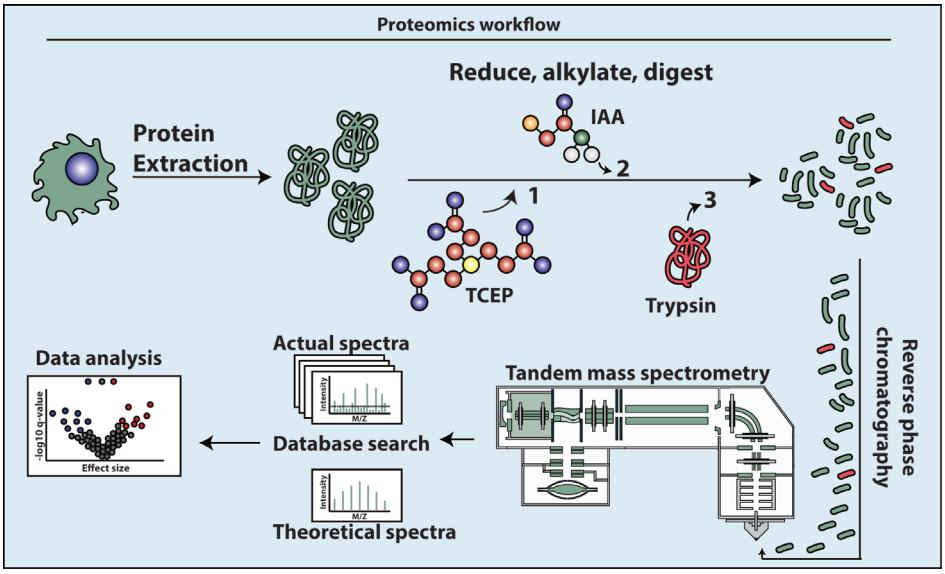



**Figure 3:** The label-free proteomics sample preparation workflow.

Briefly, total proteins are extracted from cells and the disulfide bonds are reduced and irreversibly alkylated. This creates a stable, soluble, unfolded protein which exposes all the amino acids to a protease. The irreversible alkylation of the S-H groups found in cysteine with iodoacetamide causes the formation of carboxyamidomethylcysteine which in turn causes a cysteine mass shift from 103.01 Da to 160.03 Da. This mass shift has to be accounted for and included in the database search algorithms. Reduction and alkylation is followed by digestion of the linear proteins to peptides by a protease. In the majority of cases, trypsin can be used which will proteolytically cleave at lysine and arginine. If needed, additional enzymes can be added to increase sequencing efficiency. The peptides are separated by reverse phase chromatography on C18-silica to reduce the complexity of samples that enter the mass spectrometer. As the Thermo Scientific range of mass spectrometers were used in this study, the rest of this overview will be in line with their machines. It is important to note that other vendors may offer other options and other technologies in this space. Molecules that are able to ionise will enter the mass spectrometer filtered for neutral as they pass past the active beam guide to reduce noise. Ions enter the quadrupole mass filter for precursor ion selection. Selected ions are moved to the ion routing multipole, which performs high energy collisional dissociation to fragment precursor ions. These fragmented ions are focussed in the C-trap and sent to the orbitrap mass analyser for high resolution identification. This process is repeated every 3 seconds until the chromatography has concluded. This outputs mass/charge data that is matched to theoretical spectra which has been generated from a six frame translation of a reference genome. The peptide false discovery rates are computed and proteins are assembled following the principles of parsimony. This yields expression data which can be compared using statistical techniques depending on the type of experiment done.

Mass spectrometry-based proteomics identifies and assigns abundance on an atomic basis. This allows for powerful labelling techniques which greatly increase both the accuracy and the types of experiments that can be performed. Label-free proteomics is commonly used for differential expression as modern data analytics techniques along with high resolution mass spectrometry provide adequate accuracy for analysis. In this technique two conditions are analysed separately and their profiles are combined in the data analysis steps. The number and complexity of each processing step detailed in Figure 4 increases the technical variability in the experimental setup which decreases the accuracy (Figure 4). However, label-free experiments require no special reagents and yield the greatest number of identifications compared to labelled techniques. Chemical labelling of the peptides can be done using tandem mass tags or iTraq labels. These tags are either added at the protein level to find protease sites or at the peptide level for highly accurate quantification. This is achieved by decreasing variability but

requires additional fragmentation for optimal sequencing which decreases the number of proteins that can be identified through increase in cycle times (Figure 4). Finally metabolic labelling is the most accurate as the proteomes of organisms are labelled prior to experimentation. Thus the majority of variability is biological as all samples are processed together. Additional techniques such as protein turnover and protein synthesis can be tracked using metabolic labels as well. This approach increases the complexity of the sample analysed and halves the identification rate. Dynamic exclusion will prevent repeated sequencing of the same precursor, but with SILAC labels, there are two separate precursors for the same proteins.

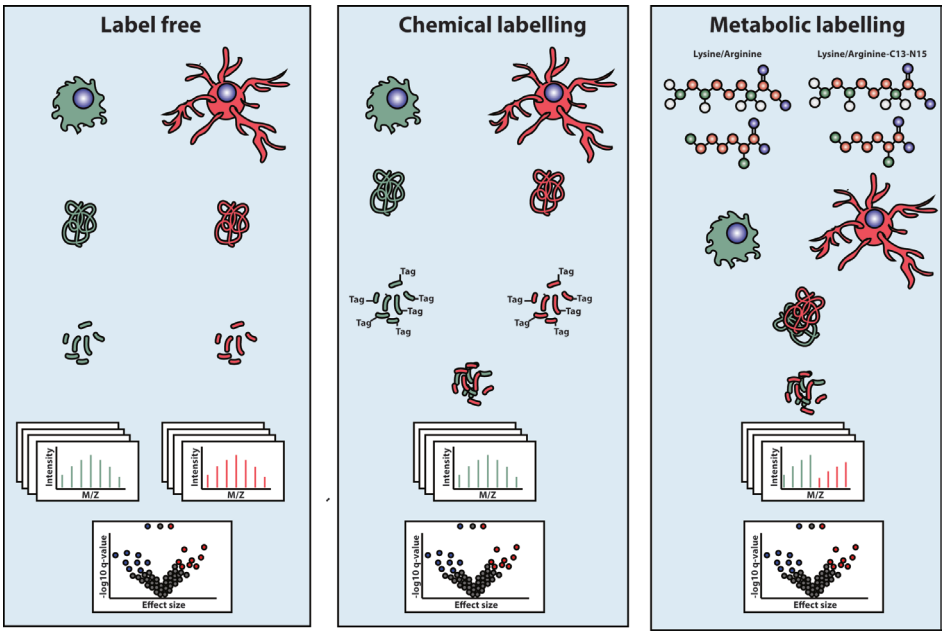



**Figure 4:** Techniques and basic workflow available in quantitative mass spectrometry namely, label free (**left**), chemical labelling with isobaric tags **(middle)** and metabolic labelling with heavy lysine and arginine (**right).**

Most studies utilising proteomics are concerned with detection as well as abundance of proteins. However, many more advanced processes can be studies using proteomics as the core technology. These include localisation studies such as localisation of organelle proteins by isotope tagging (LOPIT), detecting proteases using terminal amine isotopic labelling of substrates (TAILS), post-translational modifications (PTM), protein turnover, complex formation and more. Proteomics is thus a powerful and rapidly expanding technique which is likely to form the basis of more technologies in the near future.

## Proteogenomics

*If the genome is a list of all the instruments then the proteome is the orchestra.-R. Simpson.*

This quote elegantly described how genomics and proteomics is utilised by cells in their system wide biology. It also illustrates that one is dependent upon the other. Proteogenomics combines genome sequencing and proteomics to gain the maximum amount of information from a cell. Specifically, proteogenomics is often used to identify hidden and unique features in a non-model organism's genome and proteome which would be unachievable using either genomics or proteomics alone. These include novel proteins and protein isoforms, allele abundance, organism-specific mutations and gene fusions (Figure 5). Proteogenomics is studied by sequencing a target organism and identifying variation from the reference or assembling a new consensus genome. The open reading frames are found and translated in six frames to create an organism-specific proteome. This proteome is used as the template for theoretical spectra during mass spectrometry database searches. It is also possible to identify the variants as compared to the reference and concatenate the new altered amino sequence to an existing protein reference. The altered sequences can be retrieved with a unique ID for database searches without and will not significantly affecting the false discovery rate (40).

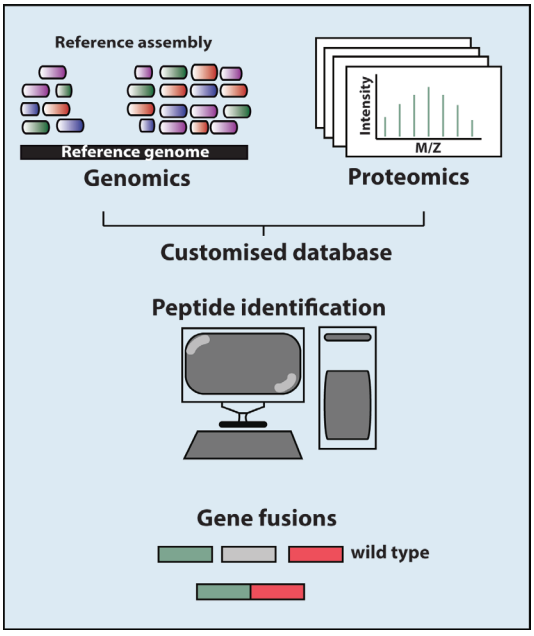



**Figure 5:** Combination of genomics and proteomics can be done computationally in order to detect unique or novel genetic elements such as gene fusions.

The omics technologies are clearly powerful and can be used in creative ways to investigate the systems biology of an organism that has not been possible until recently.

## EVOLUTION OF MYCOBACTERIUM TUBERCULOSIS.

The genus mycobacteria consists of approximately 170 species of which most are saprophytic (41). A number of these saprophytic mycobacteria can cause disease under optimal circumstances such as an immunocompromised host, but are either free-living microbes or adapted to a specific animal host, such as *M. marinum, M. abscessus* or *M. avium*. These are typically referred to as non-tuberculosis mycobacteria (NTM's) and distinct from the human-adapted mycobacterial species known as the *M. tuberculosis* complex (MTBC). The MTBC contains the major human pathogens such as *M. tuberculosis* and *M. leprae* which have no external reservoir and are dependent on a host for survival and propagation. These mycobacteria likely arose through clonal expansion of a smooth tubercle bacilli progenitor (42). The MTBC can be further sub-divided into the human-adapted strains and the animal-adapted strains, where the animal adapted strains likely branched off from *M. africanum.*

### Transition from environmental saprophyte to human pathogen

The evolution of soil or free-living mycobacteria to those dependent on mammalian hosts may have been a sequential phenomenon. A current hypothesis proposed by Neyrolles *et al* and discussed further by Gagneux *et al* is the emergence of pathogenic mycobacteria through predation of ancestral mycobacteria by free-living protozoa (43). In this scenario, organisms such as amoeba would feed on bacteria, however certain bacteria would survive and utilise the intracellular environment as a source of nutrients. This has been indicated by experimentation where multiple NTM strains are able to not only survive internalisation by *Acanthamoeba,* but actively multiply and increase in virulence with each passage (44,45). It is further suggested that through the sharing of genetic material between free-living and mycobacteria within a transition state, the ability to survive internalisation by these protozoa was transferred. Indeed a genomic island found in *M. avium* is responsible for the ability of this mycobacteria to survive within human macrophages as well as *Acanthamoeba* (46). This specific genomic island is unique to *M. avium* and likely originates from environmental proteobacteria (44).

It is tempting to speculate that a chain of events could occur where early ancestors to pathogenic mycobacteria were preyed upon by protozoa and through horizontal gene transfer acquired the ability to survive predation and thrive within the intracellular environment. The more these bacteria multiplied, the more they adapted to an

intracellular lifestyle and eventually gained the ability to survive within more complex organisms such as fish (*M. marinum*), birds *(M. avium)*, amphibians *(M. xenopi*) and mammals (*M. tuberculosis, M. leprae, M. bovis*).

## Absence of horizontal gene transfer in the MTBC

While the sharing of DNA between individuals may have given rise to the pathogenic mycobacteria, this ability was not maintained in the MTBC. The lack of horizontal gene transfer has important consequences for the evolution of bacteria. This is illustrated by the difference between vertical (parent to child) evolution and lateral (horizontal gene transfer) evolution (Figure 6). Vertical asexual evolution, as occurs in bacteria, is the movement of DNA through cloning from parent to child. Thus any changes that occur are on a small scale through mutation of the genome. Alternatively, horizontal gene transfer is the movement of DNA between organisms and is facilitated by either transformation, conjugation or transduction (Figure 6). This allows organisms within a niche to share DNA and thus functionality in its entirety. Through a combination of vertical evolution and horizontal gene transfer, bacterial organisms can acquire new genetic traits in a *de novo* fashion while remaining nimble in a dynamic environment

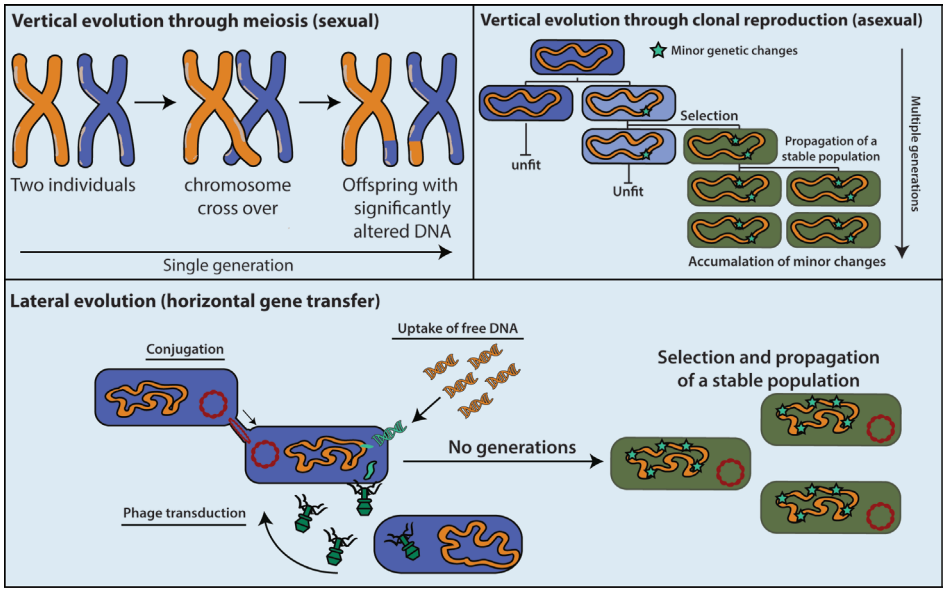



**Figure 6:** Evolution through reproduction. Vertical evolution transfers genetic information from a parent to child. This can occur through meiosis and chromosome cross over in eukaryotes (**top left)** or through clonal reproduction as seen in prokaryotes (**top right**). Genetic material can also be shared between organisms and is named horizontal gene transfer (**bottom**). Horizontal gene transfer is mediated through three main processes, namely conjugation, phage transduction or the uptake of free DNA.

through lateral DNA sharing. The acquisition of a number of bacterial traits have been attributed to horizontal gene transfer. Examples include traits such as antibiotic resistance (47), novel metabolic pathways (48), pathogenesis (49), translation (50) and more. DNA sharing is an important mechanism that introduces a significant amount of variation into the genome and is a cornerstone of evolution in the absence of sexual reproduction. Indeed the MTBC has undergone a clonal expansion to give rise to multiple strains of *M. tuberculosis*. In turn, these strains have further diversified in a restricted capacity to give rise to clades which are restricted by geographical location or adapted to animals. Interestingly, the lack of horizontal gene transfer is only found within the MTBC and the NTM's are still able to share DNA between different species. In *M. tuberculosis* the ESX-1 type VII secretion system has a major role in pathogenesis and the ESX-4 system is likely inactive. However the same systems are used for horizontal gene transfer between *M. smegmatis* and genetically distinct organisms (51) as well as receiving DNA (52). It is thus clear that at least some of the horizontal gene transfer elements have been repurposed for pathogenesis in the MTBC while still active in NTMs.

Without horizontal gene transfer, the mechanisms available to the MTBC to evolve are limited to parent to child evolution, which is less efficient when there is no crossing of chromosomes as seen in sexual reproduction. Effectively, evolution in the MTBC is limited to the DNA currently present within the bacteria. Thus changes can occur through mutations such as single nucleotide polymorphisms, deletions or duplications but acquiring novel genes and therefore novel functions are either not possible or the process surrounding it is not currently known. This is solved by horizontal gene transfer in other organisms. Nevertheless, *M. tuberculosis* is capable of surviving the host response and has seemingly adapted to the host directly as evidenced by the geographic restrictions of different clades primarily detected through conserved deletions. Single nucleotide polymorphisms also play a prominent role in the evolution of *M. tuberculosis*, specifically in the acquisition of drug resistance (53–56), through the use of an error prone polymerase (57). In this thesis we also discuss the use of gene fusions as an additional mechanisms of evolution which allows for the *de novo* biogenesis of genes in the absence of horizontal gene transfer (58).

The lack of horizontal gene transfer in *M. tuberculosis* is a fascinating topic with important implications to the continued evolution of *M. tuberculosis*. Indeed the loss of horizontal gene transfer likely resulted as a transition from NTM to MTBC where there are little to no other bacteria present in the same niche and bacterial processes can be compensated for by the host. This is seen in intracellular symbiotes, such as mitochondria, where the bacteria loses most of its genome and functionality and provides the host with a dedicated process. Perhaps the loss of horizontal gene transfer in the

MTBC would not have occurred if these organisms were not obligate pathogens, and are therefore not as restricted in their evolution or pathogenicity as it may seem.

## BACTERIAL CELL WALLS

Bacterial species can be divided into two broad categories based on their cell wall, namely Gram-positive and Gram-negative bacteria. This nomenclature is given based on the so-called Gram stain developed by Hans Christian Gram. Through the Gram stain, the physical properties of a bacterial cell wall can be described. Gram-positive bacteria will retain crystal violet dye due to a thick peptidoglycan layer. Gram-negative bacteria do not retain the dye, due to the thin peptidoglycan layer, and are only visualized by a pink counterstain (59). The Gram-positive cell wall contains an inner phospholipid membrane that divides the cytosol from a thin periplasmic layer. The thick peptidoglycan layer also contains teichoic and lipoteichoic acids that act as chelating agents and are further connected by D-D transpeptidase to form a rigid cell wall (Figure 7A, top left). In contrast to the Gram-positive cell wall, the Gram-negative cell wall has two lipid bilayers. The inner membrane is followed by a large periplasmic space which contains the thin peptidoglycan layer. This is followed by the outer membrane which has lipopolysaccharides on the surface (Figure 7, bottom). The mycobacteria

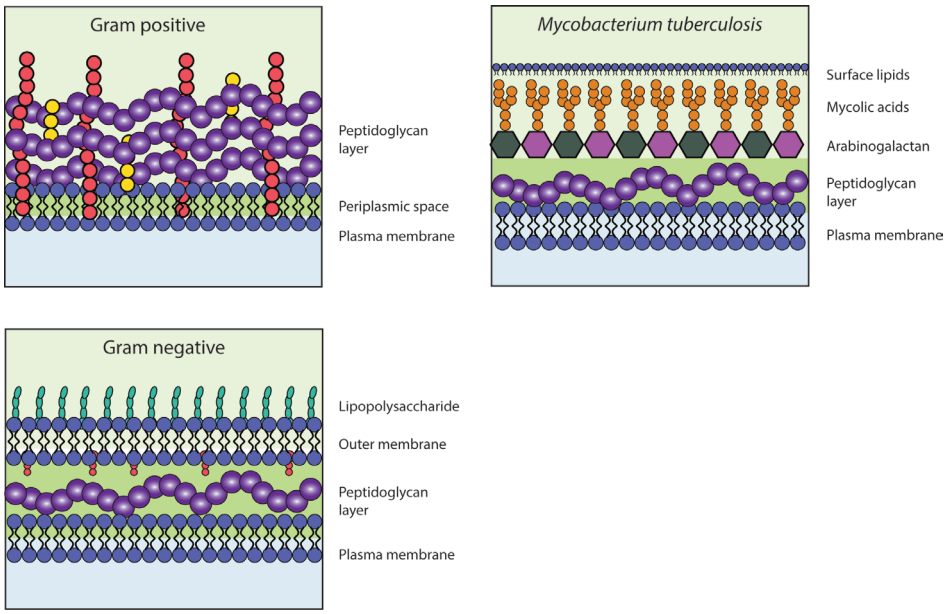



**Figure 7:** Composition of bacterial cell walls, including Gram-positive (**top left**), Mycobacterial cell wall (**top right)** and Gram-negative (**bottom**).

phylogenetically belong to the high GC Gram-positive bacteria, however these bacteria have characteristics belonging to the Gram-negative bacteria. The *M. tuberculosis* cell envelope has a cytoplasmic phospholipid membrane as well as a thin peptidoglycan layer, similar to the Gram-negatives. This is followed by a unique arabinogalactan layer and a mycolic acids layer, which is capped by surface lipids and a capsule like structure (Figure 7, top right). Because the composition of the second membrane is unique and because mycobacteria phylogenetically belong to high GC Gram-positive bacteria we have to conclude that the evolution of this second membrane was independent and therefore an example of convergent evolution.

From the cell wall composition, it becomes apparent that the needs of the Gram-positive and Gram-negative bacteria for protein secretion are radically different. Specifically, the Gram-positive bacteria only have one hydrophobic phospholipid membrane and thus has a lower barrier for protein secretion. The Gram-negative bacteria have a dual layer phospholipid membrane which unsurprisingly requires specialised systems to secrete proteins to the outer membrane. The mycobacteria have a diderm cell envelope and therefore require a specialized protein secretion system. As the inner membrane is present in all living organisms, the systems in place to secrete proteins across this membrane is highly conserved and shared across all organisms, including *M. tuberculosis*. These are namely the general secretory pathway (Sec) and twin arginine pathway (Tat) and are responsible for secreting unfolded and (partially) folded proteins, respectively (Figure 8A). The Sec system can be further split into the SecB-dependent pathway, the signal recognition particle (SRP) pathway and in some organisms the SecA2 pathway, where the SecB-dependent pathway secretes proteins to the periplasm and the SRP pathway deposits proteins in the plasma membrane (Figure 8B).

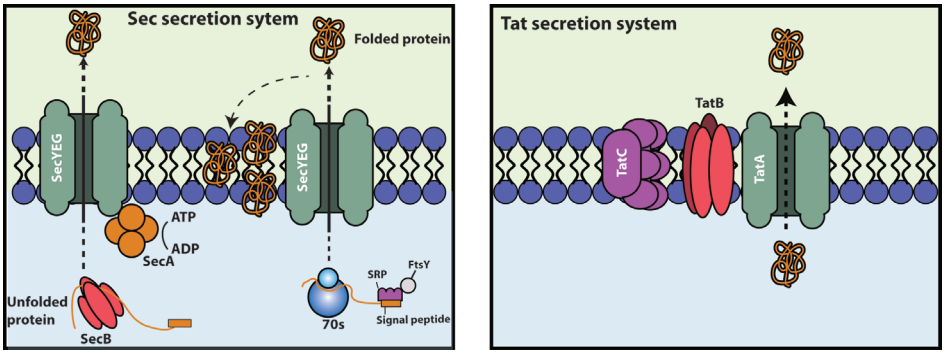



**Figure 8:** Schematic representation and basic functionality of the conserved sec **(left)** and tat **(right)** secretion systems. The sec and tat system facilitates transport across the inner cell membrane if two membranes are present. The sec system secretes unfolded proteins while the tat system is able to secrete folded proteins.

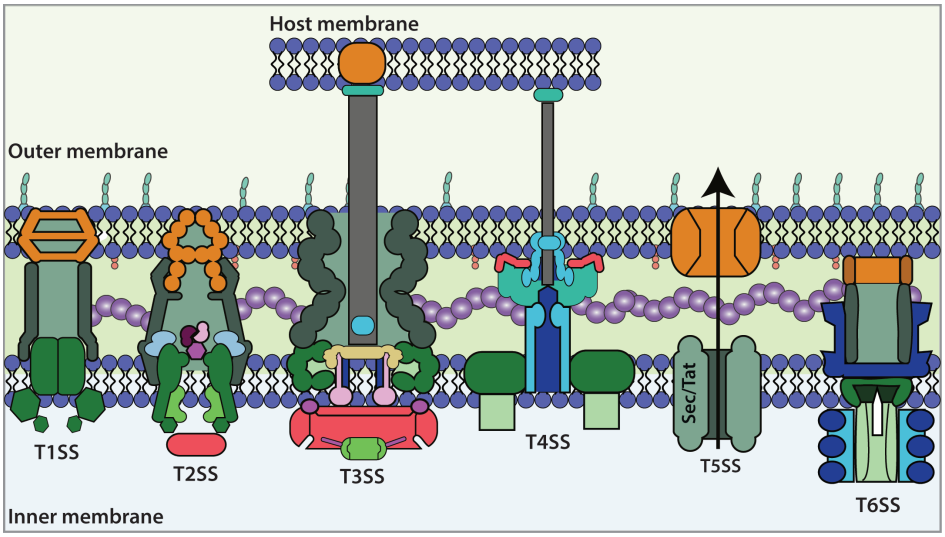



**Figure 9:** Composition and basic functionality of the secretion systems in the Gram-negative bacteria. The secretion systems are designated type I secretion system (T1SS) to type IX secretion system (T9SS). T1SS, T2SS, T3SS, T4SS and T9SS spans the inner and outer membrane.

As mentioned, Gram-negative bacteria require specialised secretion systems to transport proteins to the outer membrane. There are currently 9 known secretion systems found in the Gram-negative bacteria to suit this purpose, named Type I secretion system (T1SS) to type IX secretion system (T9SS) with the T4SS found in both Gram-positive and negative bacteria (60). Of these systems, the majority spans the inner and outer membrane, the exception is the T5SS which is only found on the outer membrane (Figure 9). Both the T5SS and the T2SS rely on the sec/tat secretion system to mediate secretion while the rest of the systems utilise a single step mechanism. Each of the systems are clearly structurally distinct, reflective of their function in the cell. Finally the T3SS and T4SS are able to interact with another cell by spanning the inner and outer membrane and injecting substrates into a host cell. T8SS forms the extracellular nucleation-precipitation pathway and is suggested to be involved in the production of adhesive fibres on the surface of pathogenic bacteria. These fibres are likely used for interaction with the host or biofilm formation (61,62). The T9SS also known as the Por secretion system is involved in gliding motility by secretion of cell surface motility adhesins (63). The composition of the bacterial cell wall clearly has a major effect on evolution of the secretion systems based on the physical requirements. Multiple specialised secretion systems evolved in the Gram-negative bacteria to compensate for the secondary outer membrane. The mycobacterial cell wall is unique, as it does not fall within the Gram-positive or negative category. The cell wall is best described as a Gram-positive cell wall with some Gram-negative traits such as an outer cell envelope

(64). The mycobacteria therefore require specialised systems to facilitate the movement of proteins across the membrane. While outer membrane proteins are present in the mycobacteria, the method by which proteins are secreted to the extracellular milieu is still unknown. The type VII secretion system is a specialised secretion system first identified in the mycobacteria which may fulfil this role and is further discussed below.

## Type VII secretion systems

The type VII secretion system was first discovered in the mycobacteria and after this discovery, variants of this system have been identified in other Gram-positive bacteria as well (65). This secretion system consists of a maximum of five loci and this number varies by species. The loci are named ESX1 to ESX-5 and are clonal expansions of each other where ESX-4 was likely the original system and ESX-2 and ESX-5 the most recent (66). This results in each system sharing specific core components as well as variable components across the homologs in the same genome. Each type VII loci is thus capable of a different function in the mycobacteria, depending on the species as well as the system (Figure 10).

The core components are comprised of the ESX conserved components (Ecc) ranging from EccA to EccD and MycP, with EccE present in all ESX systems except ESX-4, and MycP (32). Other non-secretory and non-conserved components include the ESX-specific proteins (ESP) and finally secreted components which are either small proteins such as EsxA and EsxB and the PE/PPE proteins (32). The secreted proteins of this system can also be seen as the effectors of the secretion system. This genetic basis of the T7SS has provided the first picture of how the secretion machinery would operate in *M. tuberculosis.* Importantly, this secretion system facilitates secretion through the inner membrane with an unknown mechanism secreting proteins through the secondary capsule layer. To consolidate this picture of the T7SS crystal structures of the ESX-5 and ESX-3 systems have been resolved recently (67–69). Through these endeavours a clearer picture of the Type VII secretion system, based on the conserved components, has emerged. In the current view, EccC is the pore and operates the active transport of effectors across the membrane by hydrolysing ATP. This hydrolysis is mediated by ATPase domains which undergo a conformational change in response to a substrate binding and opening the pore and allowing secretion (67).

Of the five type VII secretion systems, only ESX-1, -3 and 5 have known functionality. ESX-1 is responsible for secreting the major virulence factors ESAT-6 and CFP-10 and has been implicated in multiple aspects of pathogenesis and is the subject of multiple studies (70–75). ESX-3 is involved in nutrient uptake, specifically iron and heme uptake

(76,77). The most relevant to this thesis is the ESX-5 system, which is only found within the pathogenic mycobacteria; ESX-5 is essential for growth, is involved in carbon uptake and secretes a family of unique proteins named the PE/PPE proteins (33,39,78).

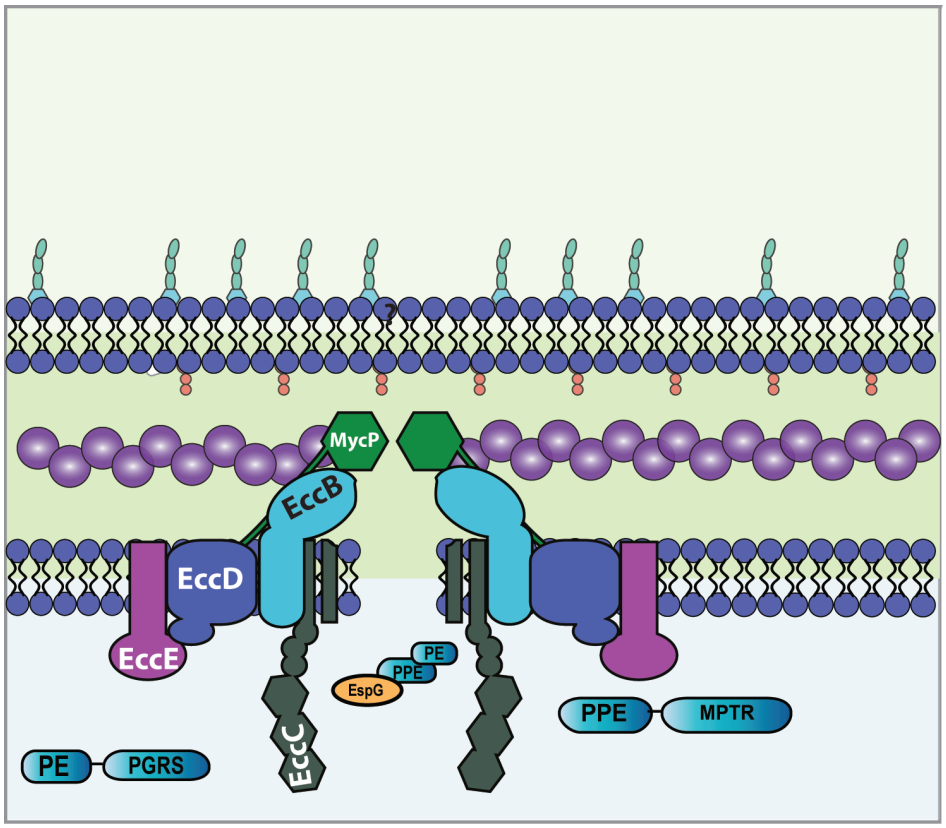



**Figure 10:** Current known structure of the T7SS. This structure is extrapolated from the T7SS found in *M. xenopi*. The T7SS is further divided into 5 different loci in the pathogenic mycobacteria named ESX-1 to ESX-5 secretion system. The ESX-5 T7SS is unique to the pathogenic mycobacteria and responsible for the secretion of PE-PGRS and PPE-MPTR proteins.

## The PE/PPE proteins

While the ESAT-6 and CFP-10 proteins are the most well-studied virulence factors of the type VII secretion system, this thesis focuses on the less well-studied PE/PPE proteins secreted by ESX-5. The PE/PPE proteins are characterised by an N-terminal containing a proline-glutamic acid motif (PE) or a proline-proline-glutamic acid motif (PPE). These protein families as a whole are found within both pathogenic and non-pathogenic mycobacteria. Members of the PE/PPE protein family can be separated into sub-groups that are restricted to either the pathogenic or non-pathogenic mycboacteria (79). The

PE/PPE proteins can be further divided into sub-families which are characterised by the C-terminal domain.

There are six possible configurations of the PE/PPE proteins, three for PE and three for PPE, where each configuration represents a sub-family. Both the PE and PPE proteins can have a unique variable C-terminal with no clear repeats. The PE proteins can also have a polymorphic GC rich sequence (PGRS) domain which is characterised by repeating GGA/GGN motifs. An N-terminal PE domain can occur as a stand-alone protein without a variable C-terminal. In addition to the unique C-terminal, PPE proteins may present with one of two types of variable C-terminal domains. The SVP C-terminal domain is characterised by the presence of a GGXSVPXXW motif. The major polymorphic tandem repeat domains (MPTR) are characterised by a repeating NXGXGNXG motif in the C-terminal (Figure 11). The secretion of both the MPTR and PGRS domains is mediated by PPE38, a SVP protein which is close to the MPTR family in sequence homology (80).

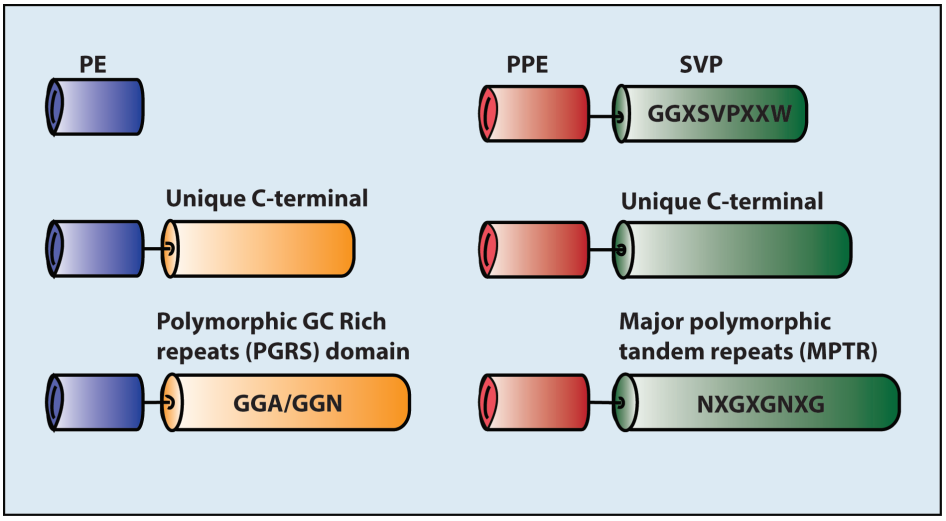



**Figure 11:** The PE/PPE proteins and the associated sub-families. The PE proteins are sub-divided into proteins that contain either a PE domain, a unique C-terminal or a PGRS domain. The PPE proteins are sub-divided into members that contain a SVP domain, a unique C-terminal or a MPTR domain. The amino acid repeats are depicted within the image. Figure adapted from *Sampson et al,* 2011 (reference 35)

The PE/PPE proteins are secreted in pairs, where a PE protein will pair with a PPE proteins, in a four helical bundle (81). The secretion signal for the PE/PPE proteins is located in the C-terminal of the PE proteins and is characterised by a YxxxD/E domain following a helix-turn-helix (82,83). This YxxxD/E domain is not found in the PPE pro-

teins, instead a WxD domain is located between the second and third alpha helix in the PE/PPE dimer (83). It has been proposed that the WxD and YxxxD/E domain interact to form a singular signal that mediates secretion through the ESX-5 system (83,84). In addition the PPE proteins has a chaperone interacting domain that binds the EspG protein associated with the specific secretion system (70,85). Together these proteins from a secretion complex that is further facilitated by PPE38 and secreted to the cell surface. It has further been shown that the PE-PGRS proteins can be cleaved from the PE domain by a novel protease, PecA. This confirms that the PGRS proteins are not only able to interact with the host cytosol in its immediate proximity but also able to circulate independently from the bacterial cell.

*M. tuberculosis* has 4173 genes of which 169 are dedicated for PE/PPE production. From these 169 genes, a total of 64 proteins encode the PE-PGRS proteins and 22 PPE-MPTR proteins. Previous computational studies has suggested that 825 proteins are secreted in *M. tuberculosis* H37Rv (86). Thus the 86 PE-PGRS and PPE-MPTR proteins equate to approximately 10% of the total secreted profile and 4% of the total coding potential. In a specialised genome, this is a large number of genes dedicated to a single protein family. These proteins are clearly important for interacting with the environment and while multiple theories have been proposed, a unified function has not yet been discovered.

## HOST PATHOGEN INTERACTIONS

Recognition and response of an infected mammalian cell to invaders is part of the innate immune response and conserved across all mammalian species. Specialised cells such as antigen presenting cells recognise pathogen associated molecular patterns (PAMPs) through specialised receptors known as pathogen recognition receptors (PRR's). In a facultative intracellular organism like *M. tuberculosis* the ESX-5 system and its effectors, the PE-PGRS and PPE-MPTR proteins, are obvious candidates as PAMPs as they are unique to pathogenic mycobacteria (79). However, by abolishing the secretion of these proteins through PPE38 or by blocking the ESX-5 system, a hypervirulence phenotype is observed (39,87).

In this thesis we also investigate the macrophage response to *ppe38* and here we will briefly introduce the most important immune signalling pathways involved. The innate immune system acts as the first barrier to invading organisms such as pathogenic bacteria or viruses. After activation of the innate immune system, the adaptive immune system will initiate, which occurs approximately four to seven after infection (88).

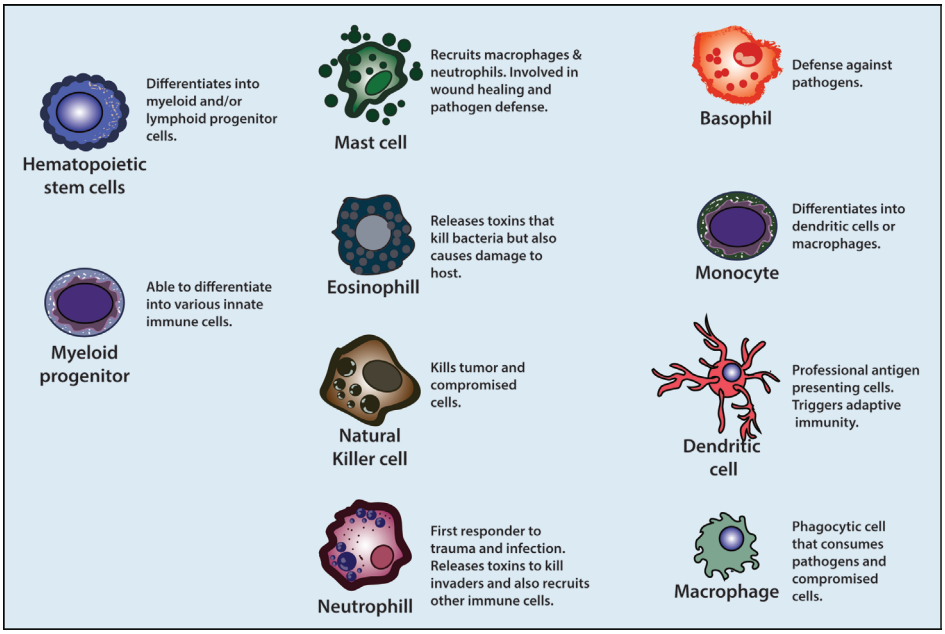



**Figure 12:** Immune cells associated with the innate immune system. Hematopoietic cells can give rise to either lymphoid (adaptive immunity) or myeloid progenitor cells (innate immunity). The myeloid cells can further differentiated into various cell types with diverse functions in combatting pathogens.

The innate immune system consists of multiple specialised cells, of which each has their own dedicated process (Figure 12) (89). After inhalation of *M. tuberculosis* the first encounter between the bacilli and the innate immune system is engulfment of bacteria by specialised alveolar macrophages (90–92). These macrophages will attempt to eliminate the invading bacilli by reactive oxygen stress, nitrogen stress, nutrient depletion and acidification (93). *M. tuberculosis* has multiple mechanisms in place to survive these initial innate immune defences. During this process inflammatory chemokines and cytokines are released which attracts additional immune cells and subsequently activates these cells (94). This inflammatory environment can overwhelm the *M. tuberculosis* bacilli in some individuals and clear the infection, in others PAMPs from the invading pathogen will be presented to naïve T-cells via antigen presenting cells, such as dendritic cells (DC), through interaction with major histocompatibility complex (MHC) (95). These T-cells can activate and initiate adaptive immune responses which can either clear or contain the bacilli. During chronic *M. tuberculosis* infection the bacilli resides mainly in DC cells in the lung interstitium (96,97). This containment can either result in the formation of dormant *M. tuberculosis,* leaving an individual asymptomatic but still infected or the bacilli can continue aggressive infection resulting in cavitation of the lung (98). In this thesis we investigate the initial response to the

PE-PGRS and PPE-MPTR proteins using a model human macrophage to stand in for alveolar macrophages. Therefore the potential signal recognition pathway and direct molecular response mediated by the macrophage in response to PE-PGRS and PPE-MPTR proteins will be introduced here.

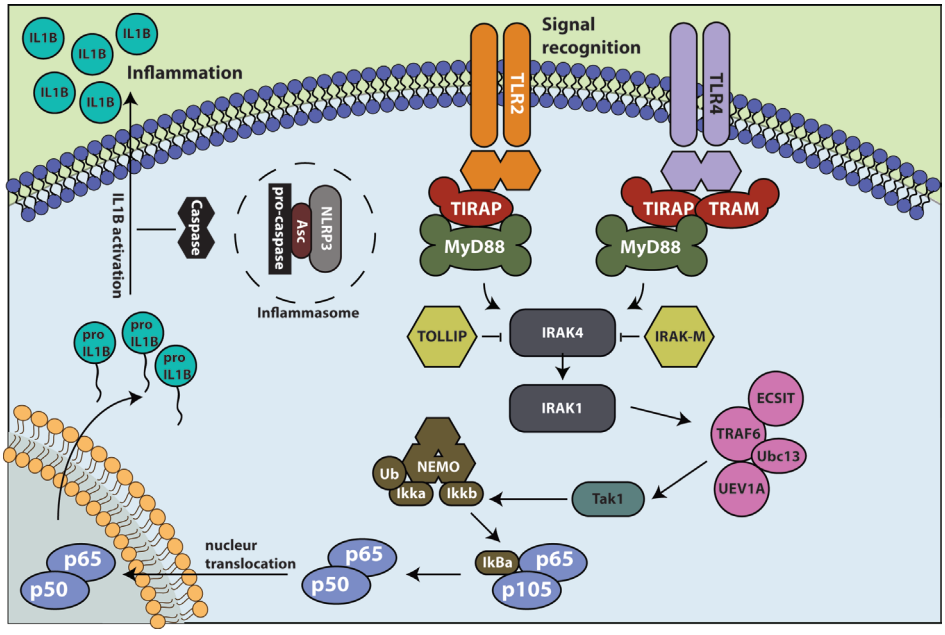



**Figure 13:** Simplified Illustration of the canonical Nuclear factor kappa B signalling pathway following either TLR2 or TLR4 MyD88 signalling pathway. This pathway results in inflammatory bursts when triggered by pathogens.

*M. tuberculosis* is recognised by the toll like receptors (TLR) present on the macrophage surface and within the cell (Figure 13). The TLR's involved are likely TLR2, TLR4 or a combination of the two (99,100). The TLRs, as well as most other pathogen recognition receptors initiate a signalling cascade that signal through the nuclear factor kappa B (NF-kB) pathway (101). This pathway controls the transcription of pro- and anti-inflammatory genes and is thus tightly regulated to avoid over and-/or under stimulation of the innate immune response on a fundamental level. This is achieved with various combinations of stimulators, inhibitors and post-translational modifications (102). Ultimately the NF-kB pathway ends in the translocation of transcription factors to the nucleus and depending on the proteins translocated different inflammatory genes can be described. There are two multimeric NF-kB proteins, NF-kB1 and NF-kB2. These proteins can either be in an inhibitory dimer state or an activated cleaved state. When cleaved, the active NF-KB sub-units, NF-kB1 p50 or NF-kB2 p52, can bind to either the RelA, RelB or c-REL and upon phosphorylation this NF-kB/Rel dimer will translocate to

the nucleus and initiate transcription of target genes. In the case of p50/RelA translocation, the canonical NF-KB pathway is stimulated and the major pro-inflammatory cytokine, Interleukin 1 beta (IL-1B), is transcribed and translated to pro-IL1B. In the pro form IL-1B is still inactive and activation occurs upon cleaving by the inflammasome. IL-1B is subsequently activated and secreted to stimulate the activation of other immune cells. This process can compound and create a cascading inflammatory response to control infection and has been reviewed in depth (102–106).

While the immune system is complex and multiple intertwined processes are in place to combat invading pathogens, it can still be overcome. How *M. tuberculosis* overcomes the immune system is still not well understood, it is however certain that the bacilli has evolved numerous strategies for this purpose.

## SCOPE OF THE THESIS

**Chapter 2** describes the testing of a variant of auxotrophic, severely attenuated mutant of *M. tuberculosis* (SAMMtb) as a substitute for virulent *M. tuberculosis* in laboratory study. This auxotroph is deficient in leucine and panthethonate biosynthesis as a result of disruption of the *leuD* and *panCD* genes. Part of this investigation involved the testing of SAMMtb in response to acidic stress using label-free proteomics. This forms the basis of our investigation into *M. tuberculosis* systems biology. We further found that SAMMtb has a greater sensitivity to cellular stress than wild type *M. tuberculosis.*

**Chapter 3** directly follows on chapter 2 where SAMMtb proteome dynamics in response to acid stress is characterised by taking advantage of the leucine auxotrophy. By exploiting the uptake of leucine, stable isotopes of leucine can be fed to the bacilli and incorporated into the proteome using a technique called stable isotope labelling of amino acid in cell culture (SILAC). Metabolic labelling has distinct advantages over label-free workflows and allows for accurate quantitative proteomics as well as advanced mass spectrometry analyses such as protein turnover and quantitative post translational modifications. We exposed SAMMtb to acidic stress over a 24h time period and characterised the protein expression, phosphorylation and protein turnover to gain a high level understanding of the *M. tuberculosis* response to common stressors. This is the first systems level analysis of a typical *M. tuberculosis* stress response that integrates multiple levels of proteome in a fully quantitative manner. Here we found that SAMMtb is more prone to enter dormancy and shutdown amino acid transport in response to acid stress. We further see an upregulation of type VII secretion components as well as effectors of the ESX-1 being dynamically phosphorylated.

**Chapter 4** investigates the evolution of *M. tuberculosis* by using genomics and proteomics in a cross-platform omics approach. The *M. tuberculosis* genome is highly specialised due to genomic decay, the lack of horizontal gene transfer and competitive co-evolution. This chapter further investigates type VII secretion and introduces the *ppe38-ppe71* mutation in *M. tuberculosis.* This mutation abolishes PE-PGRS secretion and has been found in the highly virulent lineage two isolates of *M. tuberculosis.* Initially we aimed to develop a genome processing pipeline that is able to identify deletions in the *pe/ppe* regions in clinical isolates of *M. tuberculosis* and to ultimately use this pipeline to identify the *ppe38-ppe71* deletions. In the development of this software we could resolve the breakpoints with some accuracy and noticed that open reading frames can fuse and reform proteins. We subsequently demonstrated that fused and unfused versions of the *ppe38-ppe71* mutation can be found in clinical isolates of *M. tuberculosis* and the fused genes complement PE-PGRS secretion. The protein products of these gene fusions are known as chimeric proteins and have an evolutionary implication for *M. tuberculosis*. We therefore further developed our genomics pipeline to discover more gene fusions specifically by combining reference assembly to identify multigene deletions and *de novo* assembly to find the breakpoints automatically. The consensus sequence from the *de novo* assembly was used in six frame translations and the amino acid sequences was added to a proteomics reference database if the N – and C-terminal of the chimera correspond to the two parent proteins. Using this methodology we could identify multiple known chimeras and one unknown chimeric protein. The presence of gene fusions in the mycobacterial genome is especially intriguing as there is no horizontal gene transfer. By forming gene fusions the bacilli is able to create new proteins from existing genetic material instead of relying on environmental DNA like other bacteria.

**Chapter 5** investigates the *ppe38-ppe71* deletion and the lack of PE-PGRS secretion *in vitro* and in human macrophages in an unbiased discovery based shotgun proteomics approach. In chapter 4 it became apparent that there is a vast difference between clinical isolates of *M. tuberculosis* and that each sub-lineage is different in their expression as well. Thus to accurately investigate the *ppe38-ppe71* deletion we determined the spatial protein expression patterns between wild type and mutant versions of the bacilli in a clean genetic background. We found no differences in the protein abundance in the cytoplasm, however this was not the case for the cell wall fraction and the secretomes as expected. We could identify a cluster of PE-PGRS and PPE-MPTR proteins that will likely drive the macrophage response to infection by each strain. We subsequently infected THP-1 macrophages and characterised the changes in protein abundance over time as well as the protein turnover using SILAC in response to PPE38 controlled PE-PGRS and PPE-MPTR proteins. The main finding was the dampened M1 response when macrophages were infected with *ppe38* mutants compared to wild type.

We further show that this dampening is mediated by differential translocation of the NF-kB and Rel sub-units which in turn results in altered cytokine transcription. We finally demonstrate a decrease of either IL12p70 at 48 hours post infection in the supernatant of THP-1 cells infected with either wild type and complement. Complementary to this observation, in the macrophages infected with *ppe38-ppe71* mutant we found an increase of IL-13. As we found in chapter 4, this has an interesting consequence for the virulence for lineage 2 clinical isolates, as these contain a high frequency of *ppe38-ppe71* deletions and known to mediate an anti-inflammatory response.

**Chapter 6** describes the use of quantitative proteomics to profile bronchiolar alveolar lavage (BAL) fluid from patients infected with tuberculosis. The samples obtained were from patients representing an active tuberculosis infection or the lung profile at the end of antibiotic treatment. In addition, the retention of live culturable *M. tuberculosis* was determined at the end of treatment and correlated with FDG-PET CT scans. We found that 12% of patients from a cohort of 41 retained culturable bacilli at the end of treatment and two of these patients eventually relapsed. Strikingly, within the BAL fluid proteome of the relapse patient at the end of treatment clustered closer to the active TB cohort. These types of clinical samples are rare and we could only procure one such sample. This study does however demonstrate the detection of an uncured patient by profiling the host immune response while other tests may suggest the individual is cured. With more samples and experimentation a less evasive method can be developed to screen for individuals that pose such a risk.

**Chapter 7** describes creation of a cloud-based data analysis platform for proteomics and makes the data science and statistics techniques available to the end user without the need to rely on automated systems from a service provider. During our investigation and implementation of various standard and advanced proteomics techniques, a major bottleneck was the data analysis. In order to design follow up experiments and draw conclusions it was necessary to develop multiple tools and collate them into an in-house data analysis platform. In this chapter, we explain the data cleaning and statistical quality control methods that were used in this study and the creation of a dynamic and reactive dashboard to analyse labelled and label-free proteomics data. This dashboard is freely available for use online and contains in-application descriptions of the concepts described in this chapter. The aim of this dashboard is to facilitate the speed of analysis and without the need for data science, machine learning, statistics or programming knowledge required.

In **Chapter 8** we discuss the results obtained and the concepts investigated in this thesis.

## REFERENCES


1. Hershkovitz I, Donoghue HD, Minnikin DE, May H, Lee OY-C, Feldman M, et al. Tuberculosis origin: The Neolithic scenario. Tuberculosis [Internet]. 2015 Jun [cited 2019 Jan 15];95:S122–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25726364
2. Grzybowski S, Allen EA. History and importance of scrofula. Lancet (London, England) [Internet]. 1995 Dec 2 [cited 2019 Jan 15];346(8988):1472–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7490997
3. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. J Prev Med Hyg [Internet]. 2017 Mar [cited 2019 Jan 15];58(1):E9–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28515626
4. Laennec RT. A treatise on the disease of the chest, translated by Forbes J [Internet]. New York: Hafner publishing company; 1962 [cited 2019 Jan 15]. Available from: https://archive.org/details/tuberculosistrea00klebuoft/page/n5
5. Zimmerman MR. Pulmonary and osseous tuberculosis in an Egyptian mummy. Bull New York Acad Med J Urban Heal. 1979;55(6):604–8.
6. Cave AJE, Demonstrator A. The evidence for the incidence of tuberculosis in ancient Egypt. Br J Tuberc. 1939 Jul 1;33(3):142–52.
7. The Story of Clinical Pulmonary Tuberculosis [Internet]. [cited 2020 Nov 24]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1527043/
8. The Internet Classics Archive | Of the Epidemics by Hippocrates [Internet]. [cited 2020 Nov 24]. Available from: http://classics.mit.edu/Hippocrates/epidemics.1.i.html
9. Mason PH. Spitting blood: the history of tuberculosis. Anthropol Med [Internet]. 2014 Sep 2 [cited 2020 Mar 5];21(3):357–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24963868
10. Besciu M. THE BYZANTINE PHYSICIANS. Vol. 6, Bulletin of the Transilvania University of Braşov •.
11. Eddy JJ. The ancient city of Rome, its empire, and the spread of tuberculosis in Europe. Tuberculosis. 2015 Jun 1;95(S1):S23–8.
12. Sylvius F. Opera medica [Internet]. [cited 2020 Jul 10]. Available from: https://books.google.nl/books?id=vqW3GAF2SuQC&printsec=frontcover&hl=nl&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
13. de la boe F. Sylvii Praxeos medicae idea nova, liber primus: de affectibus naturales hominis functiones laesas vel constituentibus, vel producentibus, vel consequentibus [Internet]. Elsevier. 1671 [cited 2020 Jul 10]. p. 1001. Available from: https://books.google.nl/books?id=JWYQjnS-f2UC&hl=nl&source=gbs_similarbooks
14. Roguin A. Rene theophile hyacinthe laënnec (1781-1826): The man behind the stethoscope. Vol. 4, Clinical Medicine and Research. Marshfield Clinic; 2006. p. 230–5.
15. Doetsch RN. Benjamin Marten and his "New Theory of Consumptions." Microbiol Rev [Internet]. 1978 Sep [cited 2019 Jan 15];42(3):521–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/362148
16. Daniel TMT. Jean-Antoine Villemin and the infectious nature of tuberculosis. Int J Tuberc Lung Dis [Internet]. 2015 Mar 1 [cited 2019 Jan 15];19(3):267–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC536441/

17. Budd W. MEMORANDUM ON THE NATURE AND THE MODE OF PROPAGATION OF PHTHISIS. Lancet [Internet]. 1867 Oct 12 [cited 2019 Jan 15];90(2302):451–2. Available from: https://www.sciencedirect.com/science/article/pii/S0140673602559559
18. Opal SM. A Brief History of Microbiology and Immunology. In: Vaccines: A Biography [Internet]. Springer New York; 2010 [cited 2020 Nov 24]. p. 31–56. Available from: /pmc/articles/PMC7176178/?report=abstract
19. Koch R. Die Ätiologie der Tuberkulose. [Internet]. Berlin; 1882 [cited 2019 Jan 15]. Available from: https://edoc.rki.de/bitstream/handle/176904/5163/428-445.pdf?sequence=1
20. Calmette A. L'Infection bacillaire et la tuberculose chez l'homme et chez les animaux, processus d'infection et de défense, étude biologique et expérimentale. 1922 [cited 2019 Jan 17]; Available from: https://gallica.bnf.fr/ark:/12148/bpt6k5746709s.texteImage
21. Schatz A, Bugie E, Waksman SA. Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. 1944. Clin Orthop Relat Res. 2005;437(437):3–6.
22. Comroe JH. Pay dirt - The story of Streptomycin. Part 1. From Waksman to Waksman. Am Rev Respir Dis. 1978;117(4 I):773–81.
23. Comroe JH. Pay dirt: the story of streptomycin. II. Am Rev Respir Dis. 1978;117(5):957–68.
24. FELDMAN WH, KARLSON AG, HINSHAW HC. Streptomycin in experimental tuberculosis: the effects in guinea pigs following infection in intravenous inoculation. Am Rev Tuberc [Internet]. 1947 Oct [cited 2020 Mar 5];56(4):346–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20270482
25. HOLM J. How can elimination of tuberculosis as a public health problem be achieved. Am Rev Tuberc. 1959 May 23;79(5):690–4.
26. A Timeline of HIV and AIDS | HIV.gov [Internet]. [cited 2020 Nov 24]. Available from: https://www.hiv.gov/hiv-basics/overview/history/hiv-and-aids-timeline
27. 1993 Revised Classification System for HIV Infection and Expanded Surveillance Case Definition for AIDS Among Adolescents and Adults [Internet]. [cited 2020 Mar 5]. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/00018871.htm
28. Lawn SD, Bekker L-G, Wood R. How effectively does HAART restore immune responses to Mycobacterium tuberculosis? Implications for tuberculosis control. AIDS [Internet]. 2005 Jul 22 [cited 2020 Mar 5];19(11):1113–24. Available from: http://journals.lww.com/00002030-200507220-00005
29. Suchindran S, Brouwer ES, Van Rie A. Is HIV infection a risk factor for multi-drug resistant tuberculosis? A systematic review. Vol. 4, PLoS ONE. Public Library of Science; 2009.
30. Gillespie SH. Evolution of drug resistance in Mycobacterium tuberculosis: Clinical and molecular perspective [Internet]. Vol. 46, Antimicrobial Agents and Chemotherapy. American Society for Microbiology Journals; 2002 [cited 2020 Nov 24]. p. 267–74. Available from: http://aac.asm.org/
31. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jage BB. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393(6685):537–544.
32. Bitter W, Houben ENG, Bottai D, Brodin P, Brown EJ, Cox JS, et al. Systematic genetic nomenclature for type VII secretion systems. PLoS Pathog. 2009;5(10):8–13.

33. Abdallah AM, Verboom T, Weerdenburg EM, Gey van Pittius NC, Mahasha PW, Jiménez C, et al. PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. Mol Microbiol [Internet]. 2009 Aug 1 [cited 2019 Apr 18];73(3):329–40. Available from: http://doi.wiley.com/10.1111/j.1365-2958.2009.06783.x
34. Houben ENGG, Korotkov K V., Bitter W. Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta [Internet]. 2014 Aug;1843(8):1707–16. Available from: http://dx.doi.org/10.1016/j.bbamcr.2013.11.003
35. Sampson SL. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol [Internet]. 2011;2011(Figure 1):497203. Available from: http://dx.doi.org/10.1155/2011/497203
36. L. Ramakrishnan, N. A. Federspiel and SF, Ramakrishnan L, Federspiel NA, Falkow S. Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. Science. 2000 May;288(5470):1436–1439.
37. Basu S, Pathak SK, Banerjee A, Pathak S, Bhattacharyya A, Yang Z, et al. Execution of macrophage apoptosis by PE_PGRS33 of Mycobacterium tuberculosis is mediated by toll-like receptor 2-dependent release of tumor necrosis factor-α. J Biol Chem [Internet]. 2007 Jan 12 [cited 2020 Jul 10];282(2):1039–50. Available from: http://www.jbc.org/
38. Dheenadhayalan V, Delogu G, Brennan MJ. Expression of the PE_PGRS 33 protein in Mycobacterium smegmatis triggers necrosis in macrophages and enhanced mycobacterial survival. Microbes Infect [Internet]. 2006 Jan [cited 2015 Jun 23];8(1):262–72. Available from: http://www.sciencedirect.com/science/article/pii/S1286457905002595
39. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nat Microbiol [Internet]. 2018 Feb 15 [cited 2018 Mar 6];3(2):181–8. Available from: http://www.nature.com/articles/s41564-017-0090-6
40. Heunis T, Dippenaar A, Warren RM, van Helden PD, van der Merwe RG, Gey van Pittius NC, et al. Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium tuberculosis* Isolates. J Proteome Res [Internet]. 2017 Oct 6 [cited 2018 Feb 27];16(10):3841–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28820946
41. Fedrizzi T, Meehan CJ, Grottola A, Giacobazzi E, Fregni Serpini G, Tagliazucchi S, et al. Genomic characterization of Nontuberculous Mycobacteria. Sci Rep [Internet]. 2017 Mar 27 [cited 2020 Jul 10];7. Available from: https://pubmed.ncbi.nlm.nih.gov/28345639/
42. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet [Internet]. 2013 Feb 6 [cited 2020 Jul 10];45(2):172–9. Available from: https://www.nature.com/articles/ng.2517
43. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis [Internet]. Vol. 16, Nature Reviews Microbiology. Nature Publishing Group; 2018 [cited 2020 Jul 10]. p. 202–13. Available from: https://www.nature.com/articles/nrmicro.2018.8
44. Jang J, Becq J, Gicquel B, Deschavanne P, Neyrolles O. Horizontally acquired genomic islands in the tubercle bacilli. Trends Microbiol [Internet]. 2008 Jul [cited 2020 Jul 10];16(7):303–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18515114/
45. Cirillo JD, Falkow S, Tompkins LS, Bermudez LE. Interaction of Mycobacterium avium with environmental amoebae enhances virulence. Infect Immun. 1997;65(9).
46. Danelishvili L, Wu M, Stang B, Harriff M, Cirillo S, Cirillo J, et al. Identification of Mycobacterium avium pathogenicity island important for macrophage and amoeba infection. Proc

Natl Acad Sci U S A [Internet]. 2007 Jun 26 [cited 2020 Jul 10];104(26):11038–43. Available from: www.pnas.org/cgi/content/full/

47. Von Wintersdorff CJH, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, et al. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer [Internet]. Vol. 7, Frontiers in Microbiology. Frontiers Media S.A.; 2016 [cited 2020 Jul 10]. p. 173. Available from: www.frontiersin.org
48. Marri PR, Bannantine JP, Golding GB. Comparative genomics of metabolic pathways in Mycobacterium species: Gene duplication, gene decay and lateral gene transfer. FEMS Microbiol Rev [Internet]. 2006 Nov [cited 2020 Jul 10];30(6):906–25. Available from: https://pubmed.ncbi.nlm.nih.gov/17064286/
49. Juhas M, Van Der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: Tools of bacterial horizontal gene transfer and evolution [Internet]. Vol. 33, FEMS Microbiology Reviews. Oxford Academic; 2009 [cited 2020 Jul 10]. p. 376–93. Available from: https://academic.oup.com/femsre/article-abstract/33/2/376/589749
50. Tuller T, Girshovich Y, Sella Y, Kreimer A, Freilich S, Kupiec M, et al. Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res [Internet]. 2011 Feb 22;39(11):4743–55. Available from: https://doi.org/10.1093/nar/gkr054
51. Coros A, Callahan B, Battaglioli E, Derbyshire KM. The specialized secretory apparatus ESX-1 is essential for DNA transfer in *Mycobacterium smegmatis*. Mol Microbiol [Internet]. 2008 Jul 1 [cited 2017 Oct 12];69(4):???-??? Available from: http://doi.wiley.com/10.1111/j.1365-2958.2008.06299.x
52. Gray TA, Clark RR, Boucher N, Lapierre P, Smith C, Derbyshire KM. Intercellular communication and conjugation are mediated by ESX secretion systems in mycobacteria. Science [Internet]. 2016 Oct 21 [cited 2017 Oct 12];354(6310):347–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27846571
53. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet [Internet]. 2017 Mar 1 [cited 2020 Jul 10];49(3):395–402. Available from: https://www.nature.com/articles/ng.3767
54. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet [Internet]. 2018 Feb 1 [cited 2020 Jul 10];50(2):307–16. Available from: https://www.nature.com/articles/s41588-017-0029-0
55. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat Commun [Internet]. 2019 Dec 1 [cited 2020 Jul 10];10(1):1–11. Available from: https://doi.org/10.1038/s41467-019-10110-6
56. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med [Internet]. 2015 Dec 27 [cited 2018 Apr 25];7(1):51. Available from: http://genomemedicine.com/content/7/1/51
57. Boshoff HIM, Reed MB, Barry CE, Mizrahi V. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. Cell [Internet]. 2003 Apr 18 [cited 2020 Jul 10];113(2):183–93. Available from: http://genolist.pasteur.fr/Tuberculist

58. Identification of gene fusion events in Mycobacterium tuberculosis that encode chimeric proteinsGallant J, Mouton J, Ummels R, ten Hagen-Jongman C, Kriel N, Pain A, et al. Identification of gene fusion events in Mycobacterium tuberculosis that encode chimeric proteins. NAR Genomics Bioinforma. 2020 Jun 1;2(2).
59. Beveridge TJ, Davies JA. Cellular responses of Bacillus subtilis and Escherichia coli to the Gram stain. J Bacteriol. 1983;156(2):846–58.
60. Costa TRD, Felisberto-Rodrigues C, Meir A, Prevost MS, Redzej A, Trokter M, et al. Secretion systems in Gram-negative bacteria: structural and mechanistic insights. Nat Rev Microbiol [Internet]. 2015 May 15 [cited 2018 Jan 9];13(6):343–59. Available from: http://www.nature.com/doifinder/10.1038/nrmicro3456
61. Barnhart MM, Chapman MR. Curli Biogenesis and Function. Annu Rev Microbiol [Internet]. 2006 Oct 11 [cited 2020 Jul 10];60(1):131–47. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev.micro.60.080805.142106
62. Stathopoulos C, Hendrixson DR, Thanassi DG, Hultgren SJ, St. Geme JW, Curtiss R. Secretion of virulence determinants by the general secretory pathway in Gram-negative pathogens: An evolving story [Internet]. Vol. 2, Microbes and Infection. Elsevier Masson SAS; 2000 [cited 2020 Jul 10]. p. 1061–72. Available from: https://pubmed.ncbi.nlm.nih.gov/10967286/
63. Sato K, Naito M, Yukitake H, Hirakawa H, Shoji M, McBride MJ, et al. A protein secretion system linked to bacteroidete gliding motility and pathogenesis. Proc Natl Acad Sci U S A [Internet]. 2010 Jan 5 [cited 2020 Jul 10];107(1):276–81. Available from: www.pnas.org/cgi/doi/10.1073/pnas.0912010107
64. Chiaradia L, Lefebvre C, Parra J, Marcoux J, Burlet-Schiltz O, Etienne G, et al. Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane. Sci Rep [Internet]. 2017 Dec 1 [cited 2020 Nov 24];7(1). Available from: https://pubmed.ncbi.nlm.nih.gov/28993692/
65. Bottai D, Gröschel MI, Brosch R. Type VII secretion systems in gram-positive bacteria. In: Current Topics in Microbiology and Immunology [Internet]. Springer Verlag; 2017 [cited 2020 Nov 24]. p. 235–65. Available from: https://link.springer.com/chapter/10.1007/82_2015_5015
66. Abdallah AM, Gey van Pittius NC, DiGiuseppe Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, et al. Type VII secretion — mycobacteria show the way. Nat Rev Microbiol [Internet]. 2007 Nov 1;5:883. Available from: http://dx.doi.org/10.1038/nrmicro1773
67. Beckham KSH, Ciccarelli L, Bunduc CM, Mertens HDT, Ummels R, Lugmayr W, et al. Structure of the mycobacterial ESX-5 type VII secretion system membrane complex by single-particle analysis. Nat Microbiol [Internet]. 2017;2:17047. Available from: http://www.nature.com/articles/nmicrobiol201747
68. Poweleit N, Czudnochowski N, Nakagawa R, Trinidad D, Murphy KC, Sassetti C, et al. Title: The structure of the endogenous ESX-3 secretion system. Elife [Internet]. 2019 Dec 1 [cited 2020 Nov 24];8. Available from: /pmc/articles/PMC6986878/?report=abstract
69. Bunduc CM, Fahrenkamp D, Wald J, Ummels R, Bitter W, Houben ENG, et al. Structure and dynamics of the ESX-5 type VII secretion system of Mycobacterium tuberculosis [Internet]. bioRxiv. bioRxiv; 2020 [cited 2021 Mar 8]. p. 2020.12.02.408906. Available from: https://doi.org/10.1101/2020.12.02.408906
70. Phan TH, van Leeuwen LM, Kuijl C, Ummels R, van Stempvoort G, Rubio-Canalejas A, et al. EspH is a hypervirulence factor for Mycobacterium marinum and essential for the

secretion of the ESX-1 substrates EspE and EspF. PLoS Pathog [Internet]. 2018 Aug 1 [cited 2020 Jul 10];14(8). Available from: https://pubmed.ncbi.nlm.nih.gov/30102741/

71. Ma Y, Keil V, Sun J. Characterization of Mycobacterium tuberculosis EsxA membrane insertion: roles of N- and C-terminal flexible arms and central helix-turn-helix motif. J Biol Chem [Internet]. 2015 Mar 13 [cited 2018 Jan 15];290(11):7314–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25645924
72. Van Der Wel N, Hava D, Houben D, Fluitsma D, Van Zon M, Pierson J, et al. M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. Cell [Internet]. 2007 Jun 29 [cited 2013 Feb 5];129(7):1287–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17604718
73. Recall of long-lived immunity to Mycobacterium tuberculosis infection in mice - PubMed [Internet]. [cited 2020 Nov 24]. Available from: https://pubmed.ncbi.nlm.nih.gov/7897219/
74. Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ, Morin PM, et al. The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. Proc Natl Acad Sci U S A [Internet]. 2003 Oct 14 [cited 2018 Jan 15];100(21):12420–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14557547
75. Smith J, Manoranjan J, Pan M, Bohsali A, Xu J, Liu J, et al. Evidence for pore formation in host cell membranes by ESX-1-secreted ESAT-6 and its role in Mycobacterium marinum escape from the vacuole. Infect Immun [Internet]. 2008 Dec [cited 2020 Nov 24];76(12):5478–87. Available from: https://pubmed.ncbi.nlm.nih.gov/18852239/
76. Tufariello JM, Chapman JR, Kerantzas CA, Wong K-W, Vilchèze C, Jones CM, et al. Separable roles for *Mycobacterium tuberculosis* ESX-3 effectors in iron acquisition and virulence. Proc Natl Acad Sci [Internet]. 2016;201523321. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1523321113
77. Sloan Siegrist M, Steigedal M, Ahmad R, Mehra A, Dragset MS, Schuster BM, et al. Mycobacterial Esx-3 requires multiple components for iron acquisition. MBio [Internet]. 2014 May 6 [cited 2020 Nov 24];5(3). Available from: https://pubmed.ncbi.nlm.nih.gov/24803520/
78. Ates LS, Ummels R, Commandeur S, van der Weerd R, Sparrius M, Weerdenburg E, et al. Essential Role of the ESX-5 Secretion System in Outer Membrane Permeability of Pathogenic Mycobacteria. PLOS Genet [Internet]. 2015;11(5):e1005190. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4418733&tool=pmcentrez&rendertype=abstract
79. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM, et al. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol Biol [Internet]. 2006;6(95):1471–2148. Available from: http://dx.doi.org/10.1186/1471-2148-6-95
80. McEvoy CRE, van Helden PD, Warren RM, van Pittius N, Gey van Pittius NC. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. BMC Evol Biol [Internet]. 2009 Jan;9(1):237. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-237
81. Chen X, Cheng HF, Zhou J, Chan CY, Lau KF, Tsui SKW, et al. Structural basis of the PE–PPE protein interaction in Mycobacterium tuberculosis. J Biol Chem [Internet]. 2017 Oct 13 [cited 2020 Jul 10];292(41):16880–90. Available from: /pmc/articles/PMC5641871/?report=abstract

82. DiGiuseppe Champion PA, Stanley SA, Champion MM, Brown EJ, Cox JS. C-terminal signal sequence promotes virulence factor secretion in Mycobacterium tuberculosis. Science (80- ) [Internet]. 2006 Sep 15 [cited 2020 Jul 10];313(5793):1632–6. Available from: https://science.sciencemag.org/content/313/5793/1632
83. Daleke MH, Ummels R, Bawono P, Heringa J, Vandenbroucke-Grauls CMJE, Luirink J, et al. General secretion signal for the mycobacterial type VII secretion pathway. Proc Natl Acad Sci [Internet]. 2012 Jul 10;109(28):11342–7. Available from: http://www.pnas.org/content/109/28/11342.abstract
84. Poweleit N, Czudnochowski N, Nakagawa R, Murphy K, Sassetti C, Rosenberg OS. A large inner membrane pore defines the ESX translocon. bioRxiv [Internet]. 2019 Oct 10 [cited 2020 Jul 10];800169. Available from: https://doi.org/10.1101/800169
85. Korotkova N, Freire D, Phan TH, Ummels R, Creekmore CC, Evans TJ, et al. Structure of the Mycobacterium tuberculosis type VII secretion system chaperone EspG5 in complex with PE25-PPE41 dimer. Mol Microbiol [Internet]. 2014 Oct 1 [cited 2020 Jul 10];94(2):367–82. Available from: https://pubmed.ncbi.nlm.nih.gov/25155747/
86. Vizcaíno C, Restrepo-Montoya D, Rodríguez D, Niño LF, Ocampo M, Vanegas M, et al. Computational prediction and experimental assessment of secreted/surface proteins from Mycobacterium tuberculosis H37Rv. PLoS Comput Biol [Internet]. 2010 Jun [cited 2020 Jul 10];6(6):1–14. Available from: /pmc/articles/PMC2891697/?report=abstract
87. Weerdenburg EM, Abdallah AM, Mitra S, de Punder K, van der Wel NN, Bird S, et al. ESX-5-deficient Mycobacterium marinum is hypervirulent in adult zebrafish. Cell Microbiol [Internet]. 2012 May 1;14(5):728–39. Available from: http://dx.doi.org/10.1111/j.1462-5822.2012.01755.x
88. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. Principles of innate and adaptive immunity. 2001 [cited 2020 Jul 10]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK27090/
89. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol [Internet]. 2010 Feb [cited 2020 Jul 10];125(2 SUPPL. 2):S3. Available from: /pmc/articles/PMC2923430/?report=abstract
90. Philips JA, Ernst JD. Tuberculosis Pathogenesis and Immunity. Annu Rev Pathol Mech Dis [Internet]. 2012 Feb 28 [cited 2020 Jul 10];7(1):353–84. Available from: https://pubmed.ncbi.nlm.nih.gov/22054143/
91. Srivastava S, Ernst JD, Desvignes L. Beyond macrophages: The diversity of mononuclear cells in tuberculosis. Immunol Rev [Internet]. 2014 Nov 1 [cited 2020 Jul 10];262(1):179–92. Available from: https://pubmed.ncbi.nlm.nih.gov/25319335/
92. Cohen SB, Gern BH, Delahaye JL, Adams KN, Plumlee CR, Winkler JK, et al. Alveolar Macrophages Provide an Early Mycobacterium tuberculosis Niche and Initiate Dissemination. Cell Host Microbe [Internet]. 2018 Sep 12 [cited 2020 Jul 10];24(3):439-446.e4. Available from: https://pubmed.ncbi.nlm.nih.gov/30146391/
93. Weiss G, Schaible UE. Macrophage defense mechanisms against intracellular bacteria. Immunol Rev [Internet]. 2015 Mar 1 [cited 2020 Jul 10];264(1):182–203. Available from: /pmc/articles/PMC4368383/?report=abstract
94. Cambier CJ, O'Leary SM, O'Sullivan MP, Keane J, Ramakrishnan L. Phenolic Glycolipid Facilitates Mycobacterial Escape from Microbicidal Tissue-Resident Macrophages. Immunity [Internet]. 2017 Sep 19 [cited 2020 Jul 10];47(3):552-565.e4. Available from: http://dx.doi.org/10.1016/j.immuni.2017.08.003

95. Pennock ND, White JT, Cross EW, Cheney EE, Tamburini BA, Kedl RM. T cell responses: Naïve to memory and everything in between. Am J Physiol - Adv Physiol Educ [Internet]. 2013 Dec 1 [cited 2020 Jul 10];37(4):273–83. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089090/
96. Gonzalez-Juarrero M, Orme IM. Characterization of murine lung dendritic cells infected with Mycobacterium tuberculosis. Infect Immun [Internet]. 2001 [cited 2020 Jul 10];69(2):1127–33. Available from: https://pubmed.ncbi.nlm.nih.gov/11160010/
97. Wolf AJ, Linas B, Trevejo-Nuñez GJ, Kincaid E, Tamura T, Takatsu K, et al. Mycobacterium tuberculosis Infects Dendritic Cells with High Frequency and Impairs Their Function In Vivo . J Immunol [Internet]. 2007 Aug 15 [cited 2020 Jul 10];179(4):2509–19. Available from: https://pubmed.ncbi.nlm.nih.gov/17675513/
98. Lin PL, Flynn JL. Understanding Latent Tuberculosis: A Moving Target. J Immunol [Internet]. 2010 Jul 1 [cited 2020 Jul 10];185(1):15–22. Available from: /pmc/articles/PMC3311959/?report=abstract
99. Faridgohar M, Nikoueinejad H. New findings of Toll-like receptors involved in Mycobacterium tuberculosis infection [Internet]. Vol. 111, Pathogens and Global Health. Taylor and Francis Ltd.; 2017 [cited 2020 Jul 10]. p. 256–64. Available from: /pmc/articles/PMC5560203/?report=abstract
100. Grover S, Sharma T, Singh Y, Kohli S, Manjunath P, Singh A, et al. The PGRS domain of Mycobacterium tuberculosis PE_PGRS protein Rv0297 is involved in Endoplasmic reticulum stress-mediated apoptosis through toll-like receptor 4. MBio [Internet]. 2018 May 1 [cited 2020 Jul 10];9(3). Available from: /pmc/articles/PMC6016250/?report=abstract
101. Kawai T, Akira S. Signaling to NF-κB by Toll-like receptors [Internet]. Vol. 13, Trends in Molecular Medicine. Elsevier; 2007 [cited 2020 Jul 10]. p. 460–9. Available from: www.sciencedirect.com
102. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther [Internet]. 2017 Jul 14 [cited 2019 Mar 29];2:17023. Available from: http://www.nature.com/articles/sigtrans201723
103. Sun S-C. The non-canonical NF-κB pathway in immunity and inflammation. Nat Rev Immunol [Internet]. 2017 Jun 5 [cited 2019 Mar 26];17(9):545–58. Available from: http://www.nature.com/doifinder/10.1038/nri.2017.52
104. Zhang Q, Lenardo MJ, Baltimore D. 30 Years of NF-kB: A Blossoming of Relevance to Human Pathobiology. Cell [Internet]. 2017 [cited 2019 Sep 1];168:37–57. Available from: http://dx.doi.org/10.1016/j.cell.2016.12.012
105. Cildir G, Low KC, Tergaonkar V. Noncanonical NF-κB Signaling in Health and Disease [Internet]. Vol. 22, Trends in Molecular Medicine. Elsevier Ltd; 2016 [cited 2020 Jul 10]. p. 414–29. Available from: http://www.cell.com/article/S147149141600054X/fulltext
106. Dorrington MG, Fraser IDC. NF-κB signaling in macrophages: Dynamics, crosstalk, and signal integration [Internet]. Vol. 10, Frontiers in Immunology. Frontiers Media S.A.; 2019 [cited 2020 Jul 10]. p. 705. Available from: www.frontiersin.org

# 2

# Comprehensive characterization of the attenuated double auxotroph *Mycobacterium tuberculosis ΔleuDΔpanCD* as an alternative to H37Rv

Jomien M. Mouton[1]
Tiaan Heunis[1,2]
Anzaan Dippenaar[1]
James L. Gallant[1,3]
Léanie Kleynhans[1]
Samantha L. Sampson[1]*

[1] DST/NRF Centre of Excellence for Biomedical Tuberculosis Research/South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[2] Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, United Kingdom
[3] Section of Molecular Microbiology, Amsterdam Institute of Molecules, Medicines, and Systems, Vrije Universiteit Amsterdam, The Netherlands

Frontiers in microbiology

DOI: 10.3389/fmicb.2019.01922

## ABSTRACT

Currently available model organisms such as *Mycobacterium smegmatis* and *Mycobacterium bovis* Bacillus Calmette-Guérin (BCG) have significantly contributed to our understanding of tuberculosis (TB) biology, these models have limitations such as differences in genome size, growth rates and virulence. However, attenuated *Mycobacterium tuberculosis* strains may provide more representative, safer models to study *M. tuberculosis* biology. For example, the *M. tuberculosis ΔleuDΔpanCD* double auxotroph, has undergone rigorous *in vitro* and *in vivo* safety testing. Like other auxotrophic strains, this has subsequently been approved for use in biosafety level (BSL) 2 facilities. Auxotrophic strains have been assessed as models for drug-resistant *M. tuberculosis* and for studying latent TB. These offer the potential as safe and useful models, but it is important to understand how well these recapitulate salient features of non-attenuated *M. tuberculosis*. We therefore performed a comprehensive comparison of *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*. These strains demonstrated similar *in vitro* and intra-macrophage replication rates, similar responses to anti-TB agents and whole genome sequence conservation. Shotgun proteomics analysis suggested that *M. tuberculosis ΔleuDΔpanCD* has a heightened stress response that leads to reduced bacterial replication during exposure to acid stress, which has been verified using a dual-fluorescent replication reporter assay. Importantly, infection of human peripheral blood mononuclear cells with the 2 strains elicited comparable cytokine production, demonstrating the suitability of *M. tuberculosis ΔleuDΔpanCD* for immunological assays. We provide comprehensive evidence to support the judicious use of *M. tuberculosis ΔleuDΔpanCD* as a safe and suitable model organism for *M. tuberculosis* research, without the need for a BSL3 facility.

## CONTRIBUTION TO THE FIELD

Mycobacterium tuberculosis research requires access to biosafety level 3 facilities, emphasizing the need for mycobacterial model organisms to facilitate our understanding of *M. tuberculosis* pathogenesis. Previous work demonstrated the use of auxotrophic strains as models for drug-resistant *M. tuberculosis* and for studying latent tuberculosis, however it is important to understand how well these represent non-attenuated *M. tuberculosis*. In this study, we determined the suitability of the double auxotrophic *M. tuberculosis ΔleuDΔpanCD* strain as a model for *M. tuberculosis* research by comparing it to the widely used *M. tuberculosis* H37Rv reference strain. We provide comprehensive comparative analyses between these two strains concerning *in vitro* and intra-macrophage growth, genomic features, response to anti-tuberculosis agents, proteomic response to stress and host immune response. Our results suggest that *M. tuberculosis ΔleuDΔpanCD* is a suitable and safe alternative for *M. tuberculosis* research, which can be conducted in a biosafety level 2 laboratory.

## INTRODUCTION

In order to radically reduce TB deaths and incidence by 2030, as set out by the End TB Strategy (1), there is a need for improved TB therapies and more effective ways of studying the deadly pathogen *Mycobacterium tuberculosis*. Current research challenges include restricted access to Biosafety level 3 (BSL3) facilities and the slow growth of *M. tuberculosis*. This emphasises the need for mycobacterial model systems that can facilitate our understanding of *M. tuberculosis* pathogenesis. Despite studies showing the use of currently available model organisms such as *Mycobacterium smegmatis* and *Mycobacterium bovis* Bacillus Calmette-Guérin (BCG) to have significantly contributed to the understanding of *M. tuberculosis*, these models have limitations.

Apart from being non-pathogenic, *M. smegmatis* is also substantially different from *M. tuberculosis* in terms of its larger genome size and considerably shorter doubling time (2,3). The model organism BCG is known to contain a natural RD1 deletion (4), which encodes the known virulence factors early secreted antigenic target-6 kDa (ESAT-6) (5) and culture filtrate protein-10 kDa (CFP-10) (6–9), suggesting that the immune response elicited by BCG will be altered in comparison to *M. tuberculosis*.

Several attenuated strains of *M. tuberculosis* have been developed (10–18) that have the potential to serve as model organisms to study *M. tuberculosis* biology. These include strains with a deletion in the bioA gene, disrupting biotin synthesis (17) or single mutation in the eccCa1 gene, disrupting ESX-1 type VII secretion, resulting in reduced host immune responses and immunopathology (16). To improve safety, several *M. tuberculosis* strains with two or more attenuating mutations have been developed. These include *ΔRD1ΔpanCD* (11,13), which similarly to BCG does not include the RD1 region. Doubly auxotrophic strains include *ΔlysAΔpanCD* (11,13) and *ΔleuDΔpanCD* (12). Recent work generated a drug-susceptible and drug-resistant triple auxotrophic stain of *M. tuberculosis*, which provides a safe model for studying drug-resistant *M. tuberculosis* under BSL2 conditions (11). Another interesting development is the use of a streptomycin-dependent *M. tuberculosis* strain (*M. tuberculosis* 18b) as a model of latent TB (19). These strains offer potential as research models, but need comprehensive characterisation to assess their suitability to the research question at hand.

The severely attenuated double auxotrophic *ΔleuDΔpanCD* strain of *M. tuberculosis* (12,15) offers significant advantages over *M. smegmatis* and BCG such as similar genetic background, growth and antigenicity properties to the widely used laboratory strain *M. tuberculosis* H37Rv. For example, the growth rate of the *ΔleuDΔpanCD* strain in minimal medium supplemented with both leucine and pantothenate was shown to be similar

to that of wild-type *M. tuberculosis* in minimal medium (12). Importantly, although the strain is fully attenuated, it was shown to retain immunogenicity and protective capacity in a sensitive guinea pig TB aerosol challenge model (12,15). In addition, the severely attenuated mutant of *M. tuberculosis* has undergone rigorous *in vitro* and *in vivo* safety testing, since it was originally developed as a TB vaccine candidate (12,15). The *ΔleuDΔpanCD* strain proved to be highly attenuated in the severe combined immune deficient mouse *M. tuberculosis* model (12). The safety of this strain was further supported by evidence that it does not cause disease in simian immunodeficiency virus (SIV)-co-infected Rhesus macaques (15). Collectively this data indicates that in the case of accidental infection with this strain in humans, it would be highly unlikely to cause disease, even in immune-compromised hosts. The *ΔleuDΔpanCD* strain, therefore, holds minimal risk to human health and environment and serves as an excellent alternative model organism for TB research. Currently, several international laboratories have been granted approval to work with this and similar strains, under BSL2 conditions (20,21).

We therefore aimed to compare *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* to assess the suitability of this auxotrophic strain as a model for *M. tuberculosis* research. We provide comprehensive comparative analyses between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* with regards to *in vitro* and intra-macrophage growth, genomic background, response to anti-TB agents, proteomic response to stress, and the host immune response.

## MATERIALS AND METHODS

### Bacterial strains and culture

All bacterial strains utilized in this study are listed and described in Table 1, along with relevant plasmid information. All reagents were purchased from Sigma-Aldrich unless otherwise specified. Liquid cultures of mycobacterial strains were grown in Middlebrook 7H9 supplemented with 10% oleic acid-albumin-dextrose-catalase (OADC, Becton Dickinson, New Jersey, USA), 0.2% (v/v) glycerol and 0.05% (v/v) Tween 80 (7H9-OGT), with antibiotics as required for plasmid maintenance, at $37^0$C, with shaking at 180 rpm. *M. tuberculosis ΔleuDΔpanCD* liquid cultures were additionally supplemented with 50 µg/ml leucine and 24 µg/ml pantothenate. Electro-competent mycobacteria were prepared and transformed as described by (22). Solid media cultures of mycobacteria were grown on 7H10 agar supplemented with 10% OADC, 0.5% (v/v) glycerol and antibiotics at $37^0$C and in the case of *M. tuberculosis ΔleuDΔpanCD*, 50 µg/ml leucine and 24 µg/ml pantothenate. Mycobacterial strains expressing the bacterial luciferase operon from

plasmid pMV306hsp+LuxCDABE (23) do not require an exogenous substrate to produce light. Bioluminescence was used to measure the intracellular growth of both the reference strain *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*. The number of generations were calculated based on either OD, luminescence or median fluorescence intensity as previously described (3).

**Table 1.** Plasmids and strains.

| Plasmid/strain | Description | Source |
|---|---|---|
| pTiGc | *hsp60(ribo)-turboFP635 hsp60-gfp*, $Kan^R$, episomal | Mouton *et al.* (2016) |
| pMV306hsp+LuxCDABE | bacterial luciferase operon, $Kan^R$, episomal | Andreu et al. (2010) Addgene plasmid number 26519 |
| *M. smegmatis* $mc^2155$ | non-pathogenic, fast-growing model organism | ATCC 700084 |
| *M. tuberculosis ΔleuDΔpanCD* | Double leucine and pantothenate auxotroph | Sampson *et al.* (2004) |
| *M. tuberculosis* H37Rv | *M. tuberculosis* reference strain and progenitor of *M. tuberculosis ΔleuDΔpanCD* | ATCC 27294, gift from Prof. Barry Bloom |

ATCC, American Type Culture Collection; $Hyg^R$, hygromycin resistant; $Kan^R$, kanamycin resistant; Leu, leucine; Pan, pantothenate.

For proteomic analysis, *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* were cultured, separately, in Middlebrook 7H9 broth supplemented with dextrose-catalase (DC), 0.2% (v/v) glycerol and 0.05% (v/v) Tween 80 (7H9-DC) at $37^0$C. Mycobacterial cells were harvested (4000 rpm, 10 min, $4^0$C) at an OD600 of ~0.8, and the cells were washed twice with ice-cold phosphate-buffered saline (PBS) pH 7.4. Cells were washed with PBS to remove residual media before LC-MS/MS analyses. Cells were either stored at -$20^0$C until further processing (control), or resuspended in 7H9-DC (pH 4.5) and incubated for 48 hours at $37^0$C. Acid-stressed cultures were subsequently washed twice by centrifugation at 4000 rpm for 10 min at $4^0$C with ice-cold PBS pH7.4, and the pellets were stored at -$20^0$C.

For testing the effect of acid stress on bacterial replication using a dual-fluorescent replication reporter previously developed by our group (3), bacteria were grown in 7H9-DC containing 4 mM Theophylline, to induce the expression of TurboFP635, for 7 days until an OD600 of ~0.8 before washing with PBS. The cultures were sub-cultured in 7H9-DC without Theophylline at pH 6.5 and pH 4.5 and incubated for 48 hours at $37^0$C.

## Genomic DNA extraction

Genomic DNA was extracted by pelleting 15 ml culture at OD600 of 0.8 for 10 minutes at 4000 rpm according to previously published methods (24).

## Whole genome sequencing (WGS)

Whole genome sequencing was done on an Illumina NextSeq 550 instrument (Illumina, California, USA) using a paired-end approach with ~600 base fragment sizes. One microgram of DNA was used to prepare libraries for sequencing per the manufacturer's instructions using the NEBNext Ultra DNA library preparation kit for Illumina (New England Biolabs, Massachusetts, USA).

## WGS data analyses

The Illumina paired-end reads for all isolates were analysed with open source software as described previously (25,26). Identified variants were compared between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*. Deleted regions in the genomes with respect to the *M. tuberculosis* H37Rv NC000962.3 reference genome were detected with DELLY and alignments of these regions were visually inspected (27).

## MIC determination using BACTEC MGIT 960

Minimum inhibitory concentration (MIC) determinations were performed using the semi-automated liquid culture BACTEC MGIT 960 system (Becton Dickinson) and EpiCenter software equipped with a TB eXist module for drug susceptibility testing (28). Briefly, a bacterial suspension was prepared from MGIT subcultures according to the manufacturer's instructions (BACTEC™ MGIT™ 960 System User's Manual: Becton Dickinson Document Number MA-0117) and 0.5 ml of the suspension was added to each MGIT tube supplemented with 0.8 ml of OADC and 0.1 ml of the drug (dissolved in DMSO) at a concentration range of 0.06 to 9.0 µg/ml for rifampicin or 0.0015 to 2.0 µg/ml for isoniazid. *M. tuberculosis ΔleuDΔpanCD* were supplemented with 50 µg/ml leucine and 24 µg/ml pantothenate. The MIC was determined as the lowest drug concentration that tested susceptible (less than 100 growth units by automated reading when the control vial turned positive).

## Mammalian cell culture

RAW264.7 cells (ATCC TIB-71) were cultured in Dulbecco's Modified Eagle's Medium (DMEM), supplemented with 10% heat-inactivated fetal bovine serum (FBS) at $37^0$C in 5% CO2. Cells were passaged every 2-4 days. For infections, cells were seeded at $5x10^4$ cells per well in 96 well white plates. *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* were prepared and infected into RAW264.7 macrophages as described before (3). Bacteria were added to macrophages at a 10:1 ratio, and incubated at $37^0$C in 5% CO2 for 3 hours, prior to penicillin/streptomycin treatment and subsequent washing, to remove extracellular bacteria. Infected RAW264.7 macrophages were cultured in the presence of DMEM, supplemented with 10% FBS, 50 µg/ml leucine and 24 µg/ml pantothenate to allow growth of the auxotrophic strain inside macrophages. To assess

the uptake of bacteria by macrophages, cells were lysed with sterile distilled water and pipetting, followed by colony forming unit (CFU) determination by serial dilution plating of lysates onto 7H10 agar, with leucine and pantothenate supplementation where necessary. Macrophages infected with strains expressing the bacterial luciferase operon were assessed for bioluminescence expression using a microplate reader (POLARstar Omega, BMG Labtech). Intracellular growth was monitored by measuring bioluminescence every 24 hours for 3 days. Macrophages infected with mycobacteria that do not contain the luciferase operon were included as controls to subtract background luminescence expression from all samples.

## Isolation and infection of peripheral blood mononuclear cells (PBMCs)

PBMCs were isolated from whole blood, of healthy TST negative donors (n = 12), using Ficoll-Paque PLUS (GE Healthcare Life Sciences, Massachusetts, USA) density (D>1.077 g/ml) gradient centrifugation. Informed consent was obtained from all the subjects and the study was approved by the Ethical Review Committee of the Faculty of Health Sciences at Stellenbosch University (N16/05/070). Cells were cultured in Roswell Park Memorial Institute (RPMI) media, supplemented with 10% FBS at a density of 5x105 cells per well in 48-well plates (Greiner Bio-one, Kremsmünster, Austria). PBMCs were then infected with *M. tuberculosis* H37Rv or *M. tuberculosis ΔleuDΔpanCD* at an MOI of 10:1, treated with penicillin/streptomycin, followed by washing as described above, before adding fresh RPMI, containing 10% FBS. Uninfected and Lipopolysaccharide- (LPS; 10 µg/ml) stimulated cells were included as negative and positive controls, respectively. Supernatants were collected 24 hours post infection and stored at -80°C until cytokine analysis. To assess uptake, PBMCs were lysed with sterile distilled water and pipetting, followed by serial dilution plating and CFU determination as described above.

## Quantification of cytokine and chemokine levels by multiplex bead array

A human ProcartaPlexTM Multiplex Immunoassay (Thermo Fisher Scientific, Massachusetts, USA) was used to simultaneously quantify the levels of the following analytes in the culture supernatants: interleukin (IL)-1B, IL-12p70, granulocyte monocyte stimulating factor (GM-CSF), growth-regulated oncogene (GRO)α, interferon (IFN) γ, macrophage inflammatory protein (MIP)-1α (CCL3), tumor necrosis factor (TNF)α, RANTES, stromal cell-derived factor (SDF)-1α. The assays were performed according to the manufacturer's instructions and samples were evaluated in triplicate. The cytokine concentrations were measured on a Bio-Plex platform (Bio PlexTM, Bio-Rad Laboratories, California, USA). A standard curve ranging from 227 to 8 979.52 pg/ml for IL-1B, 7.15 to 29 426.85 pg/ml for IL12-p70, 13.69 to 51 379.68 pg/ml for GM-CSF, 1.93 to 9993

pg/ml GROα, 11.67 to 49 632.42 pg/ml for IFNγ, 1.5 to 6755.47 pg/ml for MIP-1α, 7.85 to 33 043.15 pg/ml for TNFα, 0.86 to 837.85 pg/ml for RANTES and 10.85 to 34 295.44 pg/ml for SDF-1, was used in the assay. Correlation coefficients (r2>0.9) for the standard curves were determined from transformed mean fluorescent intensity values for each cytokine. Bio-Plex Manager Software, version 6.1, was used to determine the median fluorescent intensities.

## Proteomic sample preparation and LC-MS/MS analysis

Mycobacterial cells from 25 ml culture (performed in four independent replicates) were mechanically lysed and whole-cell lysates were processed for LC-MS/MS analysis using a modified version of the filter-aided sample preparation (FASP) approach (Wiśniewski et al., 2009). A total of 1 µg peptide mixture from each sample was analysed, independently, on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific), connected to a Thermo Scientific UltiMate 3000 RSLCnano System (Thermo Fisher Scientific) (Detailed information is provided in Text S1 in supplemental material).

## Flow cytometry sample preparation, acquisition and analyses

Samples were sonicated, fixed in 4% formaldehyde for 30 min and washed twice in PBS, containing 0.05% Tween as previously described (3). Samples not immediately analysed were then stored at $4^0$C. Immediately prior to flow cytometry analyses samples were pelleted, resuspended in PBS and filtered. Samples were analysed using a FACSJazz flow cytometer (Beckton Dickinson) for GFP fluorescent intensity using a 488 nm laser (530/40 filter) and TurboFP635 fluorescent intensity using a 561 laser (610/20 filter). For each sample, 30,000 events were captured and flow cytometry data were analysed using FlowJo vX.0.07r2 software. The number of generations were calculated based on fluorescence intensity data as described before (3). Generation times are expressed as mean ±SD.

## Data analysis

We used an exploratory data analysis approach for the multiplex bead array assay. Details are indicated in Text S1 in supplemental material. All tandem mass spectra were analysed using MaxQuant 1.5.5.1 (29), and searched against a customized *M. tuberculosis* proteome database. Custom database construction was performed as previously described (30) and are detailed in Text S1 in supplemental material. Exploratory data analysis and visualization were performed in the R statistical programming language (https://www.r-project.org/).

## Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXD013677. Raw genomic data for this study have been deposited in the European Nucleotide Archive (ENA) under the project accession PRJEB32340.

# RESULTS

## Next-generation sequencing reveals sequence conservation between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv

Whole genome sequencing of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv (the reference strain in use in our laboratory and progenitor of *M. tuberculosis ΔleuDΔpanCD* was used to confirm sequence conservation outside of the leuD and panCD regions. When comparing *M. tuberculosis ΔleuDΔpanCD* to our laboratory reference strain *M. tuberculosis* H37Rv, only one unique nonsynonymous variant (I131T A>G) was identified at position 392 in Rv2988 (leuC) in *M. tuberculosis ΔleuDΔpanCD* that was not found in *M. tuberculosis* H37Rv. However, no peptide covering this region was identified in either of the two strains. Importantly, the *leuC* gene is located upstream of the *leuD* deletion and no proteomic differences were observed in downstream proteins in the attenuated and wild type strains, suggesting an absence of polar effects on expression. In agreement with the findings of Ioerger *et al*, our analysis identified 33 variants in both the attenuated *M. tuberculosis ΔleuDΔpanCD* strain and *M. tuberculosis* H37Rv progenitor, with respect to the reference strain *M. tuberculosis* H37Rv, NC000962.3 (31). However, 32 of these variants were identical between the attenuated *M. tuberculosis ΔleuDΔpanCD* and the *M. tuberculosis* H37Rv progenitor strain (Supplementary Table S1). Visual inspection of the alignment confirmed the expected 1297 bp panCD locus deletion at position 4043882 – 4045179 (*M. tuberculosis* H37Rv gene Rv3602c and Rv3601c) and the leuD deletion at position 3344036 – 3344394 (*M. tuberculosis* H37Rv gene Rv2987C) in the attenuated *M. tuberculosis ΔleuDΔpanCD*, with respect to *M. tuberculosis* H37Rv, NC000962.3. At a genomic level, *M. tuberculosis ΔleuDΔpanCD* is therefore highly similar to *M. tuberculosis* H37Rv.

## The *in vitro* and intra-macrophage growth of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv are comparable.

To confirm that the growth rates of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv were similar *in vitro*, samples were taken for OD measurements at predetermined time points (Figure 1A). Strains containing the pTiGc plasmid were included

to determine whether plasmid carriage affected bacterial growth. One-way ANOVA with Tukey's multiple comparisons test revealed no differences in growth rates between these strains ($p > 0.05$). To confirm the auxotrophic nature of the *M. tuberculosis ΔleuDΔpanCD* strain (12), OD measurements taken from strains supplemented exogenously with leucine and pantothenate showed normal growth, whereas restricted growth was observed in the absence of pantothenate (Figure 1B).

We next assessed replication of *M. tuberculosis* during macrophage infection. RAW264.7 macrophages were infected with *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv strains expressing the bacterial luciferase operon. Bioluminescence measurements demonstrated no difference in the growth of these strains, suggesting similar intracellular replication of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv (Figure 1C). Pearson correlation tests revealed a statistically significantly positive correlation between generations calculated using luminescence for *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv ($r=0.975$, $N=4$, $p=0.0250$).

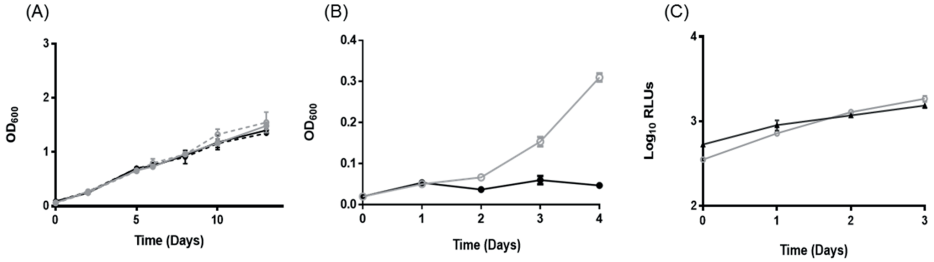



**Figure 1. Comparable growth of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv strains.** **(A)** *M. tuberculosis ΔleuDΔpanCD* (grey solid line, closed symbols), *M. tuberculosis ΔleuDΔpanCD* containing the dual reporter pTiGc (grey dotted line, open symbols), *M. tuberculosis* H37Rv (black solid line, closed symbols) and *M. tuberculosis* H37Rv containing pTiGc (black dotted line, open symbols) growth was monitored by OD. One-way ANOVA with Tukey's multiple comparisons test indicated no significant difference ($p > 0.05$). **(B)** $OD_{600}$-based *M. tuberculosis ΔleuDΔpanCD* growth with exogenous supplementation of leucine and pantothenate (grey line, open symbols) and without pantothenate supplementation (black line, closed symbols). **(C)** RAW264.7 macrophages were infected with *M. tuberculosis ΔleuDΔpanCD* + LuxCDABE (grey line, open symbols) *or M. tuberculosis* H37Rv + LuxCDABE (black line, closed symbols), and intracellular mycobacterial replication was compared by monitoring bioluminescence. Data shown are depicted as mean±SD of three technical replicates and are representative of three independent biological replicates. Pearson correlation tests revealed a statistically significantly positive correlation between generations calculated using luminescence for *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv ($r=0.975$, $N=4$, $p=0.0250$).

## *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv respond similarly to anti-tuberculosis agents

One important potential application of the attenuated *M. tuberculosis ΔleuDΔpanCD* strain is to allow for anti-mycobacterial compound screening. Phenotypic drug susceptibility testing (DST) of the two first-line anti-tuberculosis drugs with different mechanisms of action, rifampicin and isoniazid, confirmed similar minimum inhibitory concentrations (MICs) for *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv. Specifically, the MIC for isoniazid was determined to be 0.06 µg/ml for both strains, and the MIC for rifampicin was determined to be 0.25 µg/ml for *M. tuberculosis ΔleuDΔpanCD* and 0.5 µg/ml for *M. tuberculosis* H37Rv.

## Proteomic analysis reveals an increased stress response in *M. tuberculosis ΔleuDΔpanCD*.

We performed LC-MS/MS analysis on the proteomes of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv grown in media at pH 6.5 and pH 4.5. In total, we identified 21 779 unique peptides that mapped to 2 329 proteins, which contained ≥2 unique peptides at an empirical protein FDR of < 1% (Supplementary Table S2, S3). The proteins identified in this study covered 58.33% of the predicted *M. tuberculosis* H37Rv proteome. Principle component analysis revealed distinct clustering of replicates and experimental groups (Figure 2A). The first principle component (Component 1) explained 25% of the variance in the data, which has an association with acid stress. The second principle component (Component 2) explained ~19% of the variance in the data and was associated with inherent strain differences between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv. Additionally, Pearson correlation coefficients were higher within biological replicates than between groups (Figure 2B), indicating high reproducibility between independent experiments. Hierarchical clustering revealed two major clusters that separated acid stressed *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv from the strains grown under control (pH 6.5) conditions. Sub-clusters separated *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv. Taken together this data indicates that exposure to acid stress induced more variance in the data than the inherent strain differences between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv, with the strains having similar protein expression profiles under the conditions tested. However, some proteome-level differences were observed between the two strains grown under control (pH 6.5) and acid stress (pH 4.5) conditions.

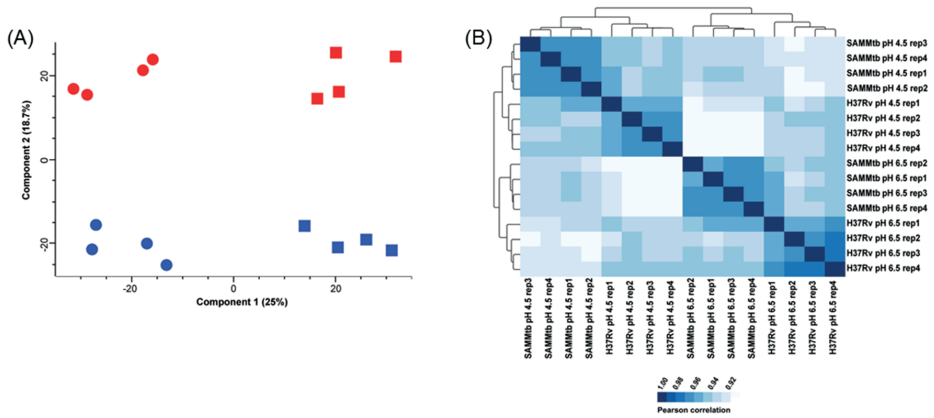



**Figure 2. Quality control analysis of proteomics data reveals distinct clustering of *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* protein intensities. (A)** Principle component analysis of z-scored MaxQuant LFQ intensities obtained from *M. tuberculosis* H37Rv (blue) and *M. tuberculosis ΔleuDΔpanCD* (red) during control (pH 6.5; square) and acid stress conditions (pH 4.5; circle). **(B)** Correlogram of Pearson correlation coefficients from MaxQuant LFQ intensities obtained for *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* (*SAMMtb*) cultured at pH 6.5 and pH 4.5.

We performed pair-wise comparisons to identify differences in relative protein abundances between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*. Twenty differentially regulated proteins, with a 2-fold change and adjusted p-value < 0.05, were identified when both of these strains were grown at pH 6.5 (Figure 3A, Supplementary Table S4). We detected 28 proteins as differentially regulated in *M. tuberculosis ΔleuDΔpanCD* during acid stress, compared to *M. tuberculosis* H37Rv (Figure 3B, Supplementary Table S5). Interestingly, we observed 8 proteins that were more abundant in acid-stressed *M. tuberculosis ΔleuDΔpanCD* cells that were also detected as more abundant in *M. tuberculosis ΔleuDΔpanCD* grown under physiological conditions, compared to *M. tuberculosis* H37Rv under the same conditions. These included proteins that play a role in dormancy, oxidative and/or nitrosative stress (AhpC, AhpD), pantothenate metabolism (CoaX, PanB, Rv3603c), transcription (Rv2989), methyl transfer (Rv2003c) and lipid catabolism (Rv2037c). This, in addition to increased abundance of DevR in *M. tuberculosis ΔleuDΔpanCD* during growth at pH 6.5, may suggest that *M. tuberculosis ΔleuDΔpanCD* experiences a more pronounced baseline stress response than *M. tuberculosis* H37Rv, which could be exacerbated during exposure to stress and could lead to reduced bacterial replication. To test this hypothesis we exploited a previously described dual-fluorescent replication reporter and flow cytometry to assess the effect of stress on bacterial replication. In accordance with our proteomics analyses, *M. tuberculosis ΔleuDΔpanCD* demonstrated a more pronounced decrease in bacterial replication in response to acid stress, compared to *M. tuberculosis* H37Rv after 48 hours (Figure 4).

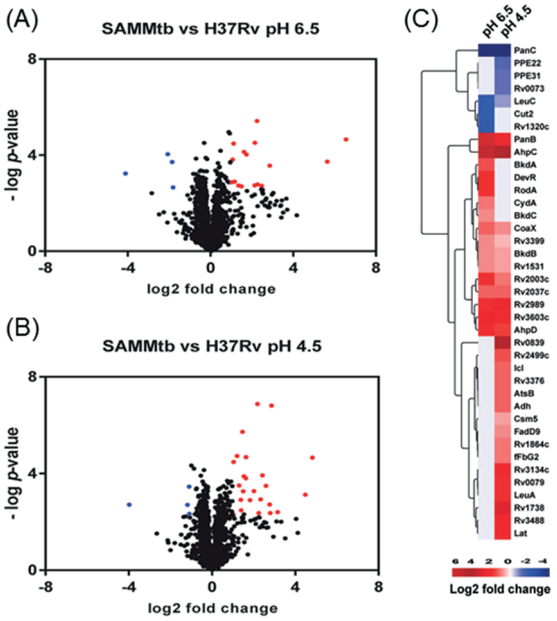



**Figure 3. Proteomic comparison of *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* under physiological *in vitro* growth conditions and in response to acid stress.** Volcano plot showing protein expression differences between *M. tuberculosis* H37Rv compared to *M. tuberculosis ΔleuDΔpanCD* when grown at pH 6.5 **(A)** and pH 4.5 **(B)**. Blue corresponds to proteins with <-1 $\log_2$ fold differential expression and adjusted p-value < 0.05. Red corresponds to proteins with >1 $\log_2$ fold differential expression and adjusted p-value < 0.05. **(C)** Heatmap of $\log_2$ fold changes of proteins showing statistically significant regulation between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* during growth at pH 6.5 and pH 4.5.

## *M. tuberculosis ΔleuDΔpanCD* induces similar or higher PBMC cytokine and chemokine responses compared to H37Rv

The concentration of 9 immune markers including cytokines, chemokines and growth factors were measured in the culture supernatant of PBMCs 24 hours after infection with *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv. No statistically significant differences were observed in RANTES, GROα, SDF-1 and IL-1B concentrations between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv (Figure 5A-D). Despite demonstrating similar trends, *M. tuberculosis ΔleuDΔpanCD* did induce higher concentrations of TNFα (p=0.0010), GM-CSF (p=0.009), MIP-1α (p=0.0021), IL-12p70 (p<0.0001) and IFNγ (p<0.0001) when compared to *M. tuberculosis* H37Rv (Figure 5E-I). *M. tuberculosis* H37Rv induced low concentrations of IL-12p70 and IFNγ. Interestingly, *M. tuberculosis ΔleuDΔpanCD* induced higher concentrations of these two cytokines (Figure 5H and I).

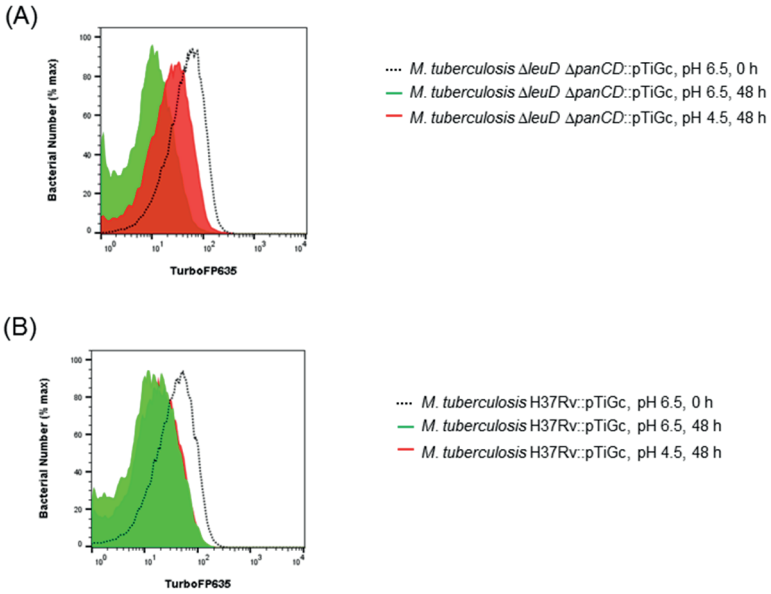



**Figure 4. Fluorescence dilution demonstrates reduced replication of *M. tuberculosis ΔleuDΔpanCD* under acid stress in comparison to *M. tuberculosis* H37Rv.** *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* containing pTiGc was cultured in the presence of 4 mM theophylline (Theo), before removal of theophylline and exposure to pH 4.5 and pH 6.5 for 48 hours prior to analyses by flow cytometry. Flow cytometry histograms demonstrate increased TurboFP635 fluorescence intensity in acidic media (pH 4.5, red), compared to normal media (pH 6.5, green) in *M. tuberculosis ΔleuDΔpanCD* (**panel A**) after 48 hours. In contrast, *M. tuberculosis* H37Rv demonstrated similar, reduced, TurboFP635 fluorescence intensity after 48 hours in acidic media (pH 4.5, red) and control media (pH 6.5, green) (**panel B**). Representative examples of three independent biological repeats are shown, *M. tuberculosis ΔleuDΔpanCD* demonstrated reduced bacterial replication (high TurboFP635 levels) in response to acid stress after 48 hours compared to *M. tuberculosis* H37R.

## DISCUSSION

We report here the assessment of attenuated *M. tuberculosis ΔleuDΔpanCD* as a suitable and safe model organism for *M. tuberculosis* research, without the need for BSL3 facilities. This strain was originally developed as a TB vaccine candidate (12,15) and deletions in the leucine and pantothenate biosynthesis pathways render it safe to work with under BSL2 conditions, since it does not replicate in the absence of exogenous supplementation with leucine and pantothenate. Importantly, we compared *M. tuberculosis ΔleuDΔpanCD* to *M. tuberculosis* H37Rv with regards to *in vitro* and intra-macrophage growth, response to anti-tuberculosis (TB) agents, genetic background, proteomic response to acid stress and host immune response. Our data supports the suitability of the attenuated strain as a model for TB research.

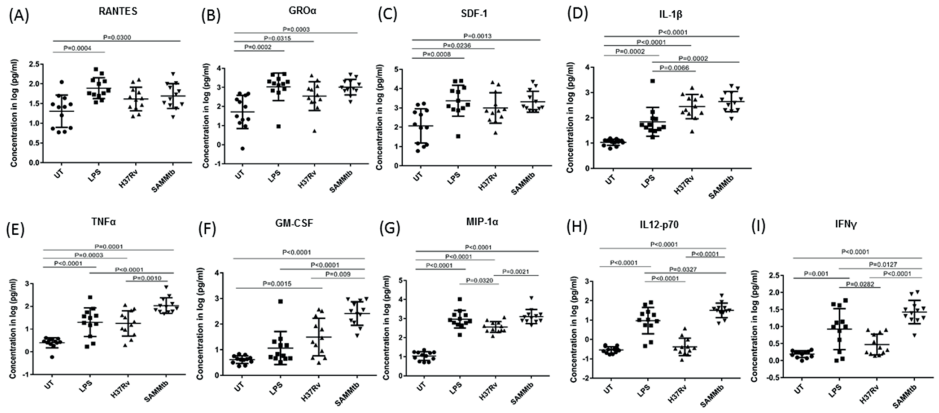



**Figure 5. Cytokine secretion in response to *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv.** Cytokine release was assessed by Luminex analyses of supernatants from PBMCs. Cells were seeded at $5x10^5$ cells per well in a 48 well plate and infected with *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv (MOI 10:1) for 24 hours. Lipopolysaccharides (LPS 10 µg/ml) and uninfected cells were included as controls. Supernatants were analysed by multiplex assays using the Bio-Plex platform. The log-transformed data were analysed using a one-way ANOVA with a Tukey Honest Significant Differences (HSD) *post hoc* test. A p-value of $< 0.05$ was regarded as significantly different. All experiments were performed in technical triplicates.

*M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD* replicated at similar rates *in vitro* and in murine macrophages; this replication was not influenced by carriage of an additional plasmid. In addition, the auxotrophic nature of *M. tuberculosis ΔleuDΔpanCD* allows for growth limitation, by removal of either pantothenate (as shown here), or leucine, or both supplements (12). This could provide a useful and tractable stress model as a complement to other commonly used dormancy models (32–36) for investigating latent TB *in vitro*.

A major goal in the field of TB drug development is shortening the course of therapy by identifying new drugs. Being able to do so without the need for a BLS3 facility would greatly decrease cost and increase accessibility to perform drug testing and screening. As proof-of-concept, we compared susceptibility to rifampicin and isoniazid, key first-line TB drugs with different mechanisms of action. Our study shows similar MICs of rifampicin and isoniazid for *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv, although a slightly increased sensitivity was observed for rifampicin for *M. tuberculosis ΔleuDΔpanCD* compared to *M. tuberculosis* H37Rv. Differences in the observed rifampicin sensitivity for *M. tuberculosis ΔleuDΔpanCD* could be attributed to the MICs being determined using serial two-fold dilutions according to the 1% proportion method; the actual MIC value may therefore be anywhere between the highest drug concentration that allows growth and the last dilution inhibiting growth. The MIC measured for

H37Rv may therefore not be 2X the MIC obtained for the attenuated auxotrophic strain, since the precision of the method is considered to be approximately 1 two-fold concentration. The observed difference in the MIC for rifampicin between the two strains is therefore within the expected variability of the assay (37,38). However, testing a wider range of anti-TB agents with different mechanisms of action would provide a more comprehensive overview of the inhibition of both strains.

Comprehensive proteomic analysis demonstrated a high level of similarity between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv when grown under physiological conditions *in vitro*. Only 20 differentially regulated proteins were identified when *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv were grown at pH 6.5, of which 16 were more abundant and four were less abundant in *M. tuberculosis ΔleuDΔpanCD*. As expected, PanC were absent in *M. tuberculosis ΔleuDΔpanCD*, which corresponds with the deletion of the *panCD* region from the attenuated *M. tuberculosis* strain (12,15).

Potentially also linked to the *panCD* deletion, we observed an increase in relative abundance of a type III pantothenate kinase (CoaX) and a 3-methyl-2-oxobutanoate hydroxymethyltransferase (PanB) that play a role in pantothenate metabolism. PanB upregulation is likely as a result of altered pantothenate metabolism introduced during construction of the *M. tuberculosis ΔleuDΔpanCD* strain. Supplementation with pantothenate in the culture medium rescues the growth defect incurred by deletion of *panCD* (as previously shown). Pantothenate is phosphorylated during Coenzyme A (CoA) biosynthesis, and CoaX could contribute to the phosphorylation of supplemented pantothenate (39).

An alkyl hydroperoxide reductase C, AhpC, and an alkyl hydroperoxide reductase, AhpD, were more abundant in *M. tuberculosis ΔleuDΔpanCD* during growth at pH6.5. AhpD reduces the active site cysteines in AhpC, an NADH-dependent thiol peroxidase, required for the detoxification of peroxides (40,41). Furthermore, a conserved protein (Rv1531) predicted to have peroxiredoxin activity was also more abundant in *M. tuberculosis ΔleuDΔpanCD*. This indicates that *M. tuberculosis ΔleuDΔpanCD* may experience increased oxidative and/or nitrosative stress during growth under physiological conditions, as compared to *M. tuberculosis* H37Rv. The transcriptional regulatory protein DevR/DosR was also more abundant in this strain, further supporting a stress response in *M. tuberculosis ΔleuDΔpanCD* (Supplementary Figure S1). To assess the possibility that the increased DevR/DosR abundance could be caused by excess clumping of the attenuated strain, we performed Ziehl-Neelsen staining of the 2 strains following culture in the presence and absence of Tween 80, as well as at pH4.5 or pH 6.5 (Supplementary Figure S2). This demonstrated no difference in clumping of the attenuated auxotrophic

strain compared to the *M. tuberculosis* H37Rv strain in any of these conditions, indicating that clumping did not influence the DevR/DosR abundance. DevR/DosR is involved in initiating the dormancy response in mycobacteria during exposure to a number of stresses (42–48). Taken together, these results indicate that the proteomic profile of *M. tuberculosis ΔleuDΔpanCD* in normal *in vitro* culture conditions largely recapitulates that of *M. tuberculosis* H37Rv. However, the attenuated strain may be skewed towards a stress response, which should be taken into account during experimental design.

We further probed the stress response of the attenuated strain by comparing proteomic profiles of *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv following exposure to acid stress. Here, 28 differentially regulated proteins were identified when *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv were exposed to pH 4.5 for 48 hours, of which 24 were more abundant and four were less abundant in *M. tuberculosis ΔleuDΔpanCD*. As expected, PanC was less abundant in acid-stressed *M. tuberculosis ΔleuDΔpanCD* when compared to acid-stressed *M. tuberculosis* H37Rv.

We observed several proteomic differences that indicate a possible increased propensity of *M. tuberculosis ΔleuDΔpanCD* to enter a heightened stress state over *M. tuberculosis* H37Rv. A dormancy-associated translation inhibitor (DATIN, Rv0079) that forms part of the DosR regulon was more abundant in acid-stressed *M. tuberculosis ΔleuDΔpanCD* (49,50). DATIN gene expression has previously been shown to be upregulated in hypoxic conditions (35) and to induce pro-inflammatory cytokine expression via interaction with Toll-like receptor 2 (51). A universal stress protein Rv3134c, also a member of the dormancy regulon, was more abundant in acid-stressed *M. tuberculosis ΔleuDΔpanCD*. This gene is the first member of the Rv3134c-devR-devS operon and has been shown to be upregulated during exposure to carbon monoxide (48,52), nitric oxide (35) and hypoxic conditions (42). Another member of the dormancy regulon, Rv1738, was also more abundant in acid-stressed *M. tuberculosis ΔleuDΔpanCD* . The Rv1738 gene has been shown to be upregulated during exposure to hypoxia (42), carbon monoxide (48) and nitric oxide (52).

LeuA is involved in leucine biosynthesis and could reflect leucine starvation in the auxotrophic *M. tuberculosis ΔleuDΔpanCD* strain under acidic conditions. It is highly likely that leucine import could be affected under acidic conditions, placing further stress (in addition to the acidic stress) on *M. tuberculosis ΔleuDΔpanCD*. Collectively, our proteomic results suggest that *M. tuberculosis ΔleuDΔpanCD* may exhibit a heightened stress response with associated metabolic changes. Specifically, exposure to an experimentally-induced stress (48 hours exposure to pH 4.5) exacerbated this stress response in *M. tuberculosis ΔleuDΔpanCD*. The *M. tuberculosis ΔleuDΔpanCD* strain can

thus serve as an ideal model to study stress responses in *M. tuberculosis* under BSL2 conditions. Furthermore, exploiting a dual-fluorescent replication reporter and flow cytometry we demonstrated markedly slower *M. tuberculosis ΔleuDΔpanCD* replication in response to acid stress after 48 hours of exposure, compared to *M. tuberculosis* H37Rv. It is thus tempting to speculate that the *M. tuberculosis ΔleuDΔpanCD* strain will enter a viable, but non-replicating ("dormant" or "persister") state more readily than *M. tuberculosis* H37Rv when exposed to unfavourable conditions. However, this hypothesis requires further validation.

Another important potential application of attenuated strains is for immunological assays. Although BCG has been widely used for this (53,54), it lacks the RD1 region and as a result it does not secrete many immunogenic proteins. Often cytokine production levels are a major concern with regards to host cells infected by attenuated strains, since many of them have essential immunogenic proteins missing. Here we show that cytokine and chemokine production by PBMCs from individuals infected with *M. tuberculosis ΔleuDΔpanCD* is not restricted and that several key cytokines (RANTES, GROα, SDF-1 and IL-1B) are produced at comparable levels by PBMCs infected with *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv. While the research question would need to be considered, *M. tuberculosis ΔleuDΔpanCD* would in many instances be a good representative strain to use as a BSL2-appropriate alternative to *M. tuberculosis* H37Rv for immunological assays. Interestingly, we observed a higher inflammatory phenotype for PBMCs infected with the attenuated *M. tuberculosis* strain in comparison to the laboratory strain, H37Rv. Specifically, TNFα, GM-CSF, MIP-1α, IL-12p70 and IFNγ were produced at higher levels by PBMCs infected with *M. tuberculosis ΔleuDΔpanCD* than those infected with *M. tuberculosis* H37Rv. IL-12 plays an important role in anti-tuberculosis cell-mediated immunity, and in addition to IL-18 are regarded as the primary inducers of IFNy production in inflammatory reactions (55–57). Several *M. tuberculosis* strains from different genetic backgrounds have demonstrated differences in the inflammatory response they elicit (58–60). More specifically, the response of human macrophages to evolutionarily modern strains (bearing the TbD1 deletion, such as Euro-American and Beijing strains) showed a lower cytokine and chemokine production compared to ancestral strains (59). Also, macrophages infected with non-Beijing strains such as Haarlem and LAM, showed heterogeneous cytokine and chemokine production compared to the Beijing strains that tend to induce homogeneously low cytokine and chemokine production. A more recent study has specifically shown that modern Beijing strains show less induction of IL1-B, IFNγ and IL-22 *in vitro*, compared to ancient Beijing and Euro-American reactivation strains (60). Despite the highly similar genetic background, the attenuated *M. tuberculosis ΔleuDΔpanCD* strain elicited higher production of analysed cytokines, akin to the more 'ancient' *M. tuberculosis* lineages.

Our proteomic analysis indicated increased production of DATIN in acid-stressed *M. tuberculosis ΔleuDΔpanCD*, which have previously been implicated in increased proinflammatory cytokine expression (51). It is therefore tempting to speculate that this may contribute to increased inflammatory responses, but this remains to be experimentally determined.

## CONCLUSION

We provide comprehensive evidence to support the judicious application of *M. tuberculosis ΔleuDΔpanCD* as a model organism for TB research. The strain recapitulates many characteristics of non-attenuated *M. tuberculosis* H37Rv, and is especially suitable for researchers interested in working with *M. tuberculosis* where access to BSL3 facilities is restricted or unavailable, or where specific instrumentation may not be available in a BSL3 setting. *M. tuberculosis ΔleuDΔpanCD* may find application in growth-based assays, drug testing, studies of dormancy/persistence, omics analysis (transcriptomics, proteomics and lipidomics), depending on research needs. As with all other models, its suitability should be carefully considered in the context of the research question. However, findings reported here can assist researchers with making an informed choice when using model organisms for tuberculosis research.

## AUTHOR CONTRIBUTIONS

JM and SS conceptualized the experiments, and drafted the manuscript. JM performed *in vitro* and intracellular growth curves, flow cytometry analyses, DST testing. TH and JG executed proteomics analyses, JG and LK contributed to luminex analyses and AD performed NGS analyses. JM, TH, AD, JG, LK and SS contributed to drafting and revising the manuscript.

## FUNDING

The authors acknowledge the SA MRC Centre for TB Research and DST/NRF Centre of Excellence for Biomedical Tuberculosis Research for financial support for this work. SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539. JL was supported by the NRF-VU Desmond Tutu Doctoral

Training Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NRF.



## CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## ACKNOWLEDGEMENTS

We thank Dr. Hanri Calitz for technical assistance with the growth curves, Mrs. Claudia Spies and Dr. Frik Sirgel for assistance with DST testing. We thank the South African Bioinformatics Initiative for advice on bioinformatics analyses of the multiplex bead array. We acknowledge the Central Analytical Facility for the use of the FACSJazz flow cytometer and Orbitrap Fusion Tribrid mass spectrometer. We thank Dr. James Posey and the Centres for Disease Control and Prevention, Atlanta, GA, USA for NGS sequencing analyses.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01922/full#supplementary-material

**Text S1:** Materials and Methods

**Table S1:** Unique and overlapping variants between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*

**Table S2:** All peptides identified in this study.

**Table S3:** All protein groups identified in this study.

**Table S4:** Regulated protein groups in *M. tuberculosis ΔleuDΔpanCD* versus *M. tuberculosis* H37Rv during two-day exposure to 7H9 at pH 6.5.

**Table S5:** Regulated protein groups in *M. tuberculosis ΔleuDΔpanCD* versus *M. tuberculosis* H37Rv during growth in 7H9 at pH 4.5.

## REFERENCES


1. GLOBAL TUBERCULOSIS REPORT 2018 [Internet]. 2018 [cited 2020 Nov 25]. Available from: http://apps.who.int/bookorders.
2. Gill WP, Harik NS, Whiddon MR, Liao RP, Mittler JE, Sherman DR. A replication clock for Mycobacterium tuberculosis. Nat Med [Internet]. 2009 Feb [cited 2020 Nov 25];15(2):211–4. Available from: https://pubmed.ncbi.nlm.nih.gov/19182798/
3. Mouton JM, Helaine S, Holden DW, Sampson SL. Elucidating population-wide mycobacterial replication dynamics at the single-cell level. Microbiol (United Kingdom). 2016;162(6):966–78.
4. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science (80- ) [Internet]. 1999;284. Available from: http://dx.doi.org/10.1126/science.284.5419.1520
5. Harboe M, Oettinger T, Wiker HG, Rosenkrands I, Andersen P. Evidence for occurrence of the ESAT-6 protein in Mycobacterium tuberculosis and virulent Mycobacterium bovis and for its absence in Mycobacterium bovis BCG. Infect Immun [Internet]. 1996 [cited 2020 Nov 25];64(1):16–22. Available from: https://pubmed.ncbi.nlm.nih.gov/8557334/
6. Lewis KN, Liao R, Guinn KM, Hickey MJ, Smith S, Behr MA, et al. Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guérin attenuation. J Infect Dis [Internet]. 2003 Jan 1 [cited 2020 Nov 25];187(1):117–23. Available from: /pmc/articles/PMC1458498/?report=abstract
7. Gao LY, Guo S, McLaughlin B, Morisaki H, Engel JN, Brown EJ. A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion. Mol Microbiol [Internet]. 2004 Sep 1 [cited 2020 Oct 30];53(6):1677–93. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2958.2004.04261.x
8. Guinn KM, Hickey MJ, Mathur SK, Zakel KL, Grotzke JE, Lewinsohn DM, et al. Individual RD1 -region genes are required for export of ESAT-6/CFP-10 and for virulence of Mycobacterium tuberculosis. Mol Microbiol [Internet]. 2004 Jan [cited 2020 Nov 25];51(2):359–70. Available from: https://pubmed.ncbi.nlm.nih.gov/14756778/
9. Fortune SM, Jaeger A, Sarracino DA, Chase MR, Sassetti CM, Sherman DR, et al. Mutually dependent secretion of proteins required for mycobacterial virulence. Proc Natl Acad Sci U S A [Internet]. 2005 Jul 26 [cited 2020 Nov 25];102(30):10676–81. Available from: https://pubmed.ncbi.nlm.nih.gov/16030141/
10. Hondalus MK, Bardarov S, Russell R, Chan J, Jacobs WR, Bloom BR. Attenuation of and protection induced by a leucine auxotroph of Mycobacterium tuberculosis. Infect Immun [Internet]. 2000 May;68(5):2888–98. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=97501&tool=pmcentrez&rendertype=abstract
11. Sambandamurthy VK, Wang X, Chen B, Russell RG, Derrick S, Collins FM, et al. A pantothenate auxotroph of Mycobacterium tuberculosis is highly attenuated and protects mice against tuberculosis. Nat Med [Internet]. 2002 [cited 2020 Nov 25];8(10):1171–4. Available from: https://pubmed.ncbi.nlm.nih.gov/12219086/
12. Sampson SL, Dascher CC, Sambandamurthy VK, Russell RG, Jacobs WR, Bloom BR, et al. Protection Elicited by a Double Leucine and Pantothenate Auxotroph of Mycobacterium tuberculosis in Guinea Pigs. Infect Immun [Internet]. 2004 May [cited 2020 Nov 24];72(5):3031–7. Available from: /pmc/articles/PMC387862/?report=abstract

13. Sambandamurthy VK, Derrick SC, Jalapathy K V., Chen B, Russell RG, Morris SL, et al. Long-term protection against tuberculosis following vaccination with a severely attenuated double lysine and pantothenate auxotroph of Mycobacterium tuberculosis. Infect Immun [Internet]. 2005 Feb [cited 2020 Nov 25];73(2):1196–203. Available from: https://pubmed.ncbi.nlm.nih.gov/15664964/
14. Derrick SC, Evering TH, Sambandamurthy VK, Jalapathy K V., Hsu T, Chen B, et al. Characterization of the protective T-cell response generated in CD4-deficient mice by a live attenuated Mycobacterium tuberculosis vaccine. Immunology [Internet]. 2007 Feb [cited 2020 Nov 25];120(2):192–206. Available from: /pmc/articles/PMC2265854/?report=abstract
15. Sampson SL, Mansfield KG, Carville A, Magee DM, Quitugua T, Howerth EW, et al. Extended safety and efficacy studies of a live attenuated double leucine and pantothenate auxotroph of Mycobacterium tuberculosis as a vaccine candidate. Vaccine [Internet]. 2011 Jun 24 [cited 2020 Nov 25];29(29–30):4839–47. Available from: https://pubmed.ncbi.nlm.nih.gov/21549795/
16. Clemmensen HS, Knudsen NPH, Rasmussen EM, Winkler J, Rosenkrands I, Ahmad A, et al. An attenuated Mycobacterium tuberculosis clinical strain with a defect in ESX-1 secretion induces minimal host immune responses and pathology. Sci Rep [Internet]. 2017;7(1):46666. Available from: https://doi.org/10.1038/srep46666
17. Kar R, Nangpal P, Mathur S, Singh S, Tyagi AK. bioA mutant of Mycobacterium tuberculosis shows severe growth defect and imparts protection against tuberculosis in guinea pigs. PLoS One [Internet]. 2017 Jun 28;12(6):e0179513. Available from: https://doi.org/10.1371/journal.pone.0179513
18. Bahal RK, Mathur S, Chauhan P, Tyagi AK. An attenuated quadruple gene mutant of Mycobacterium tuberculosis imparts protection against tuberculosis in guinea pigs. Biol Open [Internet]. 2018 [cited 2020 Nov 25];7(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29242198/
19. Zhang M, Sala C, Hartkoorn RC, Dhar N, Mendoza-Losana A, Cole ST. Streptomycin-Starved 18b, a Drug Discovery Tool for Latent Tuberculosis. Antimicrob Agents Chemother [Internet]. 2012 Nov 1;56(11):5782 LP – 5789. Available from: http://aac.asm.org/content/56/11/5782.abstract
20. Movahedzadeh F, Williams A, Clark S, Hatch G, Smith D, ten Bokum A, et al. Construction of a severely attenuated mutant of Mycobacterium tuberculosis for reducing risk to laboratory workers. Tuberculosis. 2008 Sep 1;88(5):375–81.
21. Vilchèze C, Copeland J, Keiser TL, Weisbrod T, Washington J, Jain P, et al. Rational Design of Biosafety Level 2-Approved, Multidrug-Resistant Strains of Mycobacterium tuberculosis through Nutrient Auxotrophy. Nacy CA, editor. MBio [Internet]. 2018 Jul 5;9(3):e00938-18. Available from: http://mbio.asm.org/content/9/3/e00938-18.abstract
22. Snapper SB, Melton RE, Mustafa S, Kieser T, Jr WRJ. Isolation and characterization of efficient plasmid transformation mutants of Mycobacterium smegmatis. Mol Microbiol [Internet]. 1990 [cited 2020 Nov 25];4(11):1911–9. Available from: https://pubmed.ncbi.nlm.nih.gov/2082148/
23. Andreu N, Zelmer A, Fletcher T, Elkington PT, Ward TH, Ripoll J, et al. Optimisation of Bioluminescent Reporters for Use with Mycobacteria. Doherty TM, editor. PLoS One [Internet]. 2010 May 24 [cited 2020 Nov 25];5(5):e10777. Available from: https://dx.plos.org/10.1371/journal.pone.0010777

24. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of Mycobacterium tuberculosis DNA: A question of containment. J Clin Microbiol [Internet]. 2005 Jun [cited 2020 Nov 25];43(6):2996–7. Available from: /pmc/articles/PMC1151963/?report=abstract
25. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, et al. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. BMC Genomics. 2015 Oct;16:857.
26. Dippenaar A, Parsons SDC, Sampson SL, Van Der Merwe RG, Drewe JA, Abdallah AM, et al. Whole genome sequence analysis of Mycobacterium suricattae. Tuberculosis [Internet]. 2015 Dec 1 [cited 2020 Nov 25];95(6):682–8. Available from: https://pubmed.ncbi.nlm.nih.gov/26542221/
27. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics [Internet]. 2012 Sep 15 [cited 2018 Feb 27];28(18):i333–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22962449
28. Springer B, Lucke K, Calligaris-Maibach R, Ritter C, Böttger EC. Quantitative Drug Susceptibility Testing of Mycobacterium tuberculosis by Use of MGIT 960 and EpiCenter Instrumentation. J Clin Microbiol [Internet]. 2009 Jun 1;47(6):1773 LP – 1780. Available from: http://jcm.asm.org/content/47/6/1773.abstract
29. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
30. Heunis T, Dippenaar A, Warren RM, Van Helden PD, Van Der Merwe RG, Van Pittius NCG, et al. Proteogenomic investigation of strain variation in clinical mycobacterium tuberculosis isolates. J Proteome Res [Internet]. 2017 Oct 6 [cited 2020 Oct 30];16(10):3841–51. Available from: https://pubmed.ncbi.nlm.nih.gov/28820946/
31. Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, et al. Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. J Bacteriol. 2010 Jul;192(14):3645–53.
32. Wayne LG, Hayes LG. An in vitro model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence. Infect Immun [Internet]. 1996 Jun;64(6):2062–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=174037&tool=pmcentrez&rendertype=abstract
33. Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Mol Microbiol. 2002 Feb;43(3):717–31.
34. Voskuil MI, Schnappinger D, Visconti KC, Harrell MI, Dolganov GM, Sherman DR, et al. Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. J Exp Med. 2003;198(5):705–13.
35. Voskuil MI, Visconti KC, Schoolnik GK. Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. Tuberculosis (Edinb). 2004;84(3–4):218–27.
36. Leistikow RL, Morton RA, Bartek IL, Frimpong I, Wagner K, Voskuil MI. The Mycobacterium tuberculosis DosR regulon assists in metabolic homeostasis and enables rapid recovery from nonrespiring dormancy. J Bacteriol [Internet]. 2010 Mar [cited 2020 Nov 24];192(6):1662–70. Available from: https://pubmed.ncbi.nlm.nih.gov/20023019/

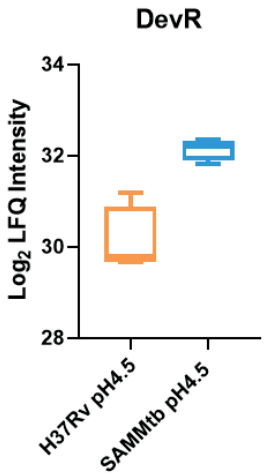
37. Kim SJ. Drug-susceptibility testing in tuberculosis: Methods and reliability of results. Eur Respir J [Internet]. 2005 Mar 1 [cited 2020 Nov 25];25(3):564–9. Available from: https://erj.ersjournals.com/content/25/3/564
38. Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis 2018.
39. Awasthy D, Ambady A, Bhat J, Sheikh G, Ravishankar S, Subbulakshmi V, et al. Essentiality and functional analysis of type I and type III pantothenate kinases of Mycobacterium tuberculosis. Microbiology [Internet]. 2010 Sep [cited 2020 Nov 25];156(9):2691–701. Available from: https://pubmed.ncbi.nlm.nih.gov/20576686/
40. Hillas PJ, Soto Del Alba F, Oyarzabal J, Wilks A, Ortiz De Montellano PR. The AhpC and AhpD antioxidant defense system of Mycobacterium tuberculosis. J Biol Chem [Internet]. 2000 Jun 23 [cited 2020 Nov 25];275(25):18801–9. Available from: https://pubmed.ncbi.nlm.nih.gov/10766746/
41. Bryk R, Lima CD, Erdjument-Bromage H, Tempst P, Nathan C. Metabolic enzymes of mycobacteria linked to antioxidant defense by a thioredoxin-like protein. Science (80- ) [Internet]. 2002 Feb 8 [cited 2020 Nov 25];295(5557):1073–7. Available from: https://pubmed.ncbi.nlm.nih.gov/11799204/
42. Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrell MI, Schoolnik GK. Regulation of the mycobacterium tuberculosis hypoxic response gene encoding α-crystallin. Proc Natl Acad Sci U S A [Internet]. 2001 Jun 19 [cited 2020 Nov 25];98(13):7534–9. Available from: www.pnas.orgcgidoi10.1073pnas.261577598
43. Boon C, Dick T. Mycobacterium bovis BCG response regulator essential for hypoxic dormancy. J Bacteriol [Internet]. 2002 Dec 15 [cited 2020 Nov 24];184(24):6760–7. Available from: http://jb.asm.org/
44. Park H-D, Guinn KM, Harrell MI, Liao R, Voskuil MI, Tompa M, et al. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis. Mol Microbiol [Internet]. 2003 Apr 9 [cited 2018 Feb 28];48(3):833–43. Available from: http://doi.wiley.com/10.1046/j.1365-2958.2003.03474.x
45. Karakousis PC, Yoshimatsu T, Lamichhane G, Woolwine SC, Nuermberger EL, Grosset J, et al. Dormancy phenotype displayed by extracellular Mycobacterium tuberculosis within artificial granulomas in mice. J Exp Med [Internet]. 2004 Sep 6 [cited 2020 Nov 25];200(5):647–57. Available from: /pmc/articles/PMC2212740/?report=abstract
46. Sharma D, Bose A, Shakila H, Das TK, Tyagi JS, Ramanathan VD. Expression of mycobacterial cell division protein, FtsZ, and dormancy proteins, DevR and Acr, within lung granulomas throughout guinea pig infection. FEMS Immunol Med Microbiol [Internet]. 2006 Dec 1 [cited 2020 Nov 25];48(3):329–36. Available from: https://academic.oup.com/femspd/article-lookup/doi/10.1111/j.1574-695X.2006.00160.x
47. Fontán P, Aris V, Ghanny S, Soteropoulos P, Smith I. Global transcriptional profile of Mycobacterium tuberculosis during THP-1 human macrophage infection. Infect Immun. 2008 Feb;76(2):717–25.
48. Kumar A, Deshane JS, Crossman DK, Bolisetty S, Yan BS, Kramnik I, et al. Heme oxygenase-1-derived carbon monoxide induces the Mycobacterium tuberculosis dormancy regulon. J Biol Chem [Internet]. 2008 Jun 27 [cited 2020 Nov 25];283(26):18032–9. Available from: /pmc/articles/PMC2440631/?report=abstract

49. Mishra S. Function prediction of rv0079, a hypothetical mycobacterium tuberculosis dosr regulon protein. J Biomol Struct Dyn [Internet]. 2009 [cited 2020 Nov 25];27(3):283–91. Available from: https://pubmed.ncbi.nlm.nih.gov/19795912/
50. Kumar A, Majid M, Kunisch R, Rani PS, Qureshi IA, Lewin A, et al. Mycobacterium tuberculosis DosR Regulon Gene Rv0079 Encodes a Putative, 'Dormancy Associated Translation Inhibitor (DATIN).' Agrewala JN, editor. PLoS One [Internet]. 2012 Jun 13 [cited 2020 Nov 25];7(6):e38709. Available from: https://dx.plos.org/10.1371/journal.pone.0038709
51. Kumar A, Lewin A, Rani PS, Qureshi IA, Devi S, Majid M, et al. Dormancy Associated Translation Inhibitor (DATIN/Rv0079) of Mycobacterium tuberculosis interacts with TLR2 and induces proinflammatory cytokine expression. Cytokine [Internet]. 2013 Oct [cited 2020 Nov 25];64(1):258–64. Available from: https://pubmed.ncbi.nlm.nih.gov/23819907/
52. Shiloh MU, Manzanillo P, Cox JS. Mycobacterium tuberculosis Senses Host-Derived Carbon Monoxide during Macrophage Infection. Cell Host Microbe [Internet]. 2008 May 15 [cited 2020 Nov 25];3(5):323–30. Available from: https://pubmed.ncbi.nlm.nih.gov/18474359/
53. Hasso-Agopsowicz M, Scriba TJ, Hanekom WA, Dockrell HM, Smith SG. Differential DNA methylation of potassium channel KCa3.1 and immune signalling pathways is associated with infant immune responses following BCG vaccination. Sci Rep. 2018 Aug;8(1):13086.
54. Whittaker E, Nicol MP, Zar HJ, Tena-Coki NG, Kampmann B. Age-related waning of immune responses to BCG in healthy children supports the need for a booster dose of BCG in TB endemic countries. Sci Rep [Internet]. 2018;8(1):15309. Available from: https://doi.org/10.1038/s41598-018-33499-4
55. Trinchieri G, Gerosa F. Immunoregulation by interleukin-12 [Internet]. Vol. 59, Journal of Leukocyte Biology. Federation of American Societies for Experimental Biology; 1996 [cited 2020 Nov 25]. p. 505–11. Available from: https://pubmed.ncbi.nlm.nih.gov/8613697/
56. Cooper AM, Magram J, Ferrante J, Orme IM. Interleukin 12 (IL-12) is crucial to the development of protective immunity in mice intravenously infected with mycobacterium tuberculosis. J Exp Med [Internet]. 1997 Jul 7 [cited 2020 Nov 25];186(1):39–45. Available from: https://pubmed.ncbi.nlm.nih.gov/9206995/
57. Dinarello CA, Novick D, Puren AJ, Fantuzzi G, Shapiro L, Mühl H, et al. Overview of interleukin-18: more than an interferon-gamma inducing factor. J Leukoc Biol. 1998 Jun;63(6):658–64.
58. Wang C, Peyron P, Mestre O, Kaplan G, van Soolingen D, Gao Q, et al. Innate immune response to Mycobacterium tuberculosis Beijing and other genotypes. PLoS One. 2010 Oct;5(10):e13594.
59. Portevin D, Gagneux S, Comas I, Young D. Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. Bessen DE, editor. PLoS Pathog [Internet]. 2011 Mar 3 [cited 2020 Nov 25];7(3):e1001307. Available from: https://dx.plos.org/10.1371/journal.ppat.1001307
60. van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LAB, et al. Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the mycobacterium tuberculosis beijing genotype. Infect Immun [Internet]. 2013 [cited 2020 Nov 25];81(10):3750–6. Available from: https://pubmed.ncbi.nlm.nih.gov/23897611/

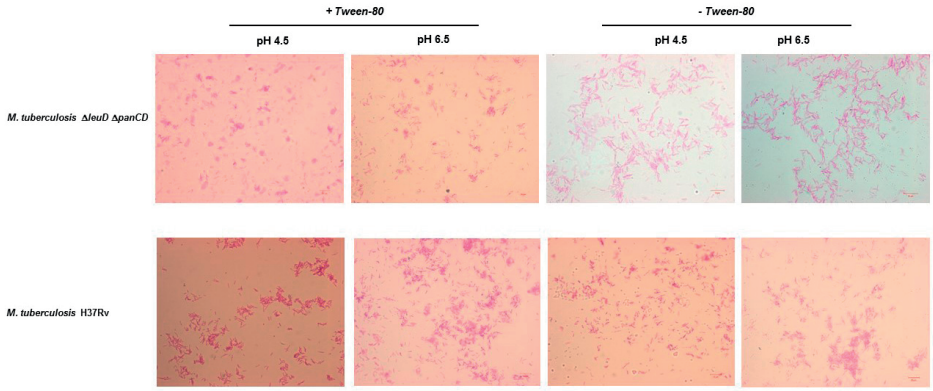
# SUPPLEMENTARY MATERIAL

**Table S1.** Unique and overlapping variants between *M. tuberculosis* H37Rv and *M. tuberculosis ΔleuDΔpanCD*

| ***M. tuberculosis* H37Rv** | | | | | | ***M. tuberculosis ΔleuDΔpanCD*** | | |
|---|---|---|---|---|---|---|---|---|
| Unique variants | | | Overlapping variants | | | Unique variants | | |
| Position | Rv number | Variant | Position | Rv number | Variant | Position | Rv number | Variant |
| | | | 14785 | Rv0012 | C233R | 3345684 | Rv2988c | I131T |
| | | | 69989 | Rv0064 | G457D | | | |
| | | | 116000 | Rv0101 | synonymous | | | |
| | | | 131174 | Intergenic | Intergenic | | | |
| | | | 390828 | Rv0323c | S142G | | | |
| | | | 459399 | Intergenic | Intergenic | | | |
| | | | 635633 | Rv0543c | synonymous | | | |
| | | | 986204 | Intergenic | Intergenic | | | |
| | | | 990001 | Rv0890c | P866A | | | |
| | | | 1010204 | Rv0907 | 1 bp insertion | | | |
| | | | 1025106 | Rv0919 | synonymous | | | |
| | | | 1037911 | Rv0930 | R305* | | | |
| | | | 1315191 | Rv1180 | *489Y | | | |
| | | | 1315884 | Rv1181 | synonymous | | | |
| | | | 1414021 | Rv1266c | R607Q | | | |
| | | | 1711627 | Rv1520 | synonymous | | | |
| | | | 2006032 | Rv1771 | Q291R | | | |
| | | | 2057774 | Rv1815 | I83F | | | |
| | | | 2207591 | Intergenic | Intergenic | | | |
| | | | 2251999 | Intergenic | Intergenic | | | |
| | | | 2282787 | Rv2037c | C312Y | | | |
| | | | 2525722 | Rv2250A | 1 bp deletion | | | |
| | | | 2718852 | Intergenic | Intergenic | | | |
| | | | 2751804 | Rv2450c | R126Q | | | |
| | | | 2809621 | Rv2495c | T107A | | | |
| | | | 2873093 | Rv2553c | 3 bp insertion | | | |
| | | | 2931837 | Rv2604c | D151E | | | |
| | | | 3580636 | Intergenic | Intergenic | | | |
| | | | 3718357 | Rv3331 | P423L | | | |
| | | | 3862472 | Intergenic | Intergenic | | | |
| | | | 3896340 | Rv3479 | L174R | | | |
| | | | 4095001 | Rv3655c | 1 bp deletion | | | |




**Supplementary figure 1: $Log_2$ LFQ intensities of DevR in *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv during acid stress.** No difference in DevR expression was observed between the two strains in acidic conditions (corrected p-value of 0.06), based on a Benjamin-Hochberg correction for multiple hypothesis testing. Differences in DevR expression levels are observed between *M. tuberculosis ΔleuDΔpanCD* and *M. tuberculosis* H37Rv during acid stress, when analyzing protein LFQ intensities without correction for multiple hypothesis testing.




**Supplementary figure 2:** Ziehl-Neelsen staining demonstrated no difference in clumping of the attenuated auxotrophic strain compared to the *M. tuberculosis* H37Rv strain in the presence and absence of Tween 80 (panel A) in media with pH 4.5 or pH 6.5 (panel B).

# 3

# Multidimensional Proteomic Analysis of *Mycobacterium tuberculosis* during Acid Stress

James L. Gallant[1,2*]
Pumla Mesatywa[1*]
Samantha L. Sampson[1]
Tiaan Heunis[1,3]

[1] DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[2] Section Molecular Microbiology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam 1081 HZ, Amsterdam, The Netherlands
[3] Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

[*]These authors contributed equally

## ABSTRACT

We previously reported on the characterisation of a leucine and pantothenate double auxotroph of *M. tuberculosis*. The leucine auxotrophy provides the opportunity to label the proteome of *M. tuberculosis* and quantitatively study various process within the proteome. We expand on the work presented previously by characterising temporal quantitative differences in protein abundance, early quantitative phosphorylation and protein homeostasis in response to acidic stress. We find that the majority of proteins are downregulated over time regardless of their importance to combat acidic stress, suggesting a metabolic shutdown. We find that the Type VII secretion system has a suggested role in the response to acid stress as well. This system is upregulated at 24 hours post stress while most other proteins are downregulated. Furthermore, the system is active at 45 minutes of acid stress as evident through increased phosphorylation and is kept in a steady state during active growth. We also find that amino acid import or utilisation is almost completely abolished in the presence of low pH, an effect that is immediately visible. Our results provide deeper understanding into the proteome response of *M. tuberculosis* to an important intracellular stress and further displays the ability of this double auxotroph to act as a model for larger quantitative *M. tuberculosis* studies.

## INTRODUCTION

*Mycobacterium tuberculosis* is the causative agent of tuberculosis (TB), a chronic granulomatous lung disease. The complexity of this disease, coupled with increased drug resistance in *M. tuberculosis*, has ensured that TB remains one of the deadliest infectious diseases worldwide. This is exemplified by the fact that 130 new TB cases per 100 000 and 1.2 million TB-related deaths were reported globally to the WHO in 2020 (1).

*M. tuberculosis* is a highly specialized pathogen with a small genome specifically evolved to survive within human macrophages. This pathogen is equipped with numerous strategies to circumvent host-derived stresses and can persist for prolonged periods of time by entering a dormant state. One of the first host responses to phagocytosed bacilli is exposure to acidic pH in the phagsomal compartment of activated macrophages (2). One of the key *M. tuberculosis* responses is to circumvent macrophage acidification by the ESX-1 secretion system (3). Thus *M. tuberculosis* located within its natural environment can experience acidic stress. This is achieved by blocking acidification which maintains the phagosome within a relatively neutral pH range (4–7). However, when the macrophages are activated by other components such as interferon gamma the phagosomal pH can drop significantly (8,9). As acid stress is clearly an important queue for the response of *M. tuberculosis,* studying the response of the bacillus can provide insights into the mechanisms employed by this pathogen to survive adverse environments. Proteomics has a number of distinct advantages over approaches such as transcriptomics for gaining insights into cellular processes. Some of these advantages include information on mature functional proteins, post-translational modification and spatial activity of the protein, all of which are lost in the genome and transcriptome. Indeed, a major consideration to note when investigating a time course response of *M. tuberculosis* is that there is a low correlation between the transcriptome and the proteome (10,11). Stable isotope labelling of amino acids in cell culture (SILAC) is a powerful method to quantify relative protein abundance with high accuracy. This technique is based on the differences of masses between the natural or “light” atomic mass compared to a “heavy” isotope which is expressed as a ratio of heavy/light. Eukaryotic cells take up essential amino acids from the environment, thus any essential amino acids can theoretically be used for SILAC-based proteomics. However, bacteria are able to synthesize amino acids naturally and more likely to follow this approach instead of relying on environmental amino acids. Deleting key steps in the amino acid biosynthesis pathway forces the bacteria to take up amino acids from the environment. Previously we have reported on an auxotrophic mutant of *M. tuberculosis* which is deficient in leucine biosynthesis (12). This provides a unique opportunity to label

*M. tuberculosis* metabolically in a BSL-2 environment. This extends the auxotrophic *M. tuberculosis* strain as a powerful model for studying *M. tuberculosis* processes.

Here we used a SILAC-based approach to study a temporal *M. tuberculosis* response to acid stress by investigation protein expression, quantitative regulation of post-translational modifications and protein turnover. By using the aforementioned approaches we performed a system-wide characterization of *M. tuberculosis* response to acid stress by using auxotrophic *M. tuberculosis* as a close genetic model. We identified increased synthesis of multiple virulence related proteins and dynamic phosphorylation of target proteins within the ESX-1 secretion system in response to acid stress. We further report a global reduction in protein synthesis and a shutdown of amino acid transport, both of which indicate a shutdown in of cellular systems and priming of the bacteria to enter a dormant state. This is the first large-scale study of *M. tuberculosis* stress response which encompasses multiple levels of regulation using fully quantitative methods. We further argue for the continued use of auxotrophic *M. tuberculosis* as a model for quantitative proteomics experiments which can be easily extended to other conditions.

## MATERIALS AND METHODS

### *M. tuberculosis* culture conditions.

*M. tuberculosis* H37Rv *ΔleuDΔpanCD* (13), a severely attenuated mutant of *M. tuberculosis* (subsequently referred to as SAMMtb), was used in this study. SAMMtb was grown in modified Sauton's broth (0.4% L-asparagine, 0.4% glucose, 0.2% citric acid, 0.05% monopotassium phosphate, 0.05% magnesium sulphate, 0.005% ferric ammonium citrate, 0.001% zinc sulphate and 0.05% Tween-80) supplemented with 50 µg/ml L-leucine, 24 µg/ml pantothenate and 50 µg/ml hygromycin-B. Cultures were maintained at 37°C without shaking. The effect of pH on SAMMtb growth was assessed by monitoring growth at pH 7.0, 5.0 and 4.5 using optical density ($OD_{600nm}$) measurements at predetermined intervals. All cultures were inoculated at an $OD_{600nm}$ of 0.05 during this study, except where stated otherwise.

### *M. tuberculosis* culturing conditions for stable isotope labelling by amino acids in cell culture.

SAMMtb was cultured in modified Sauton's broth as previously described, containing either 50 µg/ml "light" L-leucine (Leu0) or 50 µg/ml "heavy" L-leucine-5,5,5-d3 (LeuD3). SAMMtb was cultured for five weeks, with weekly sub-cultures in fresh media, to allow incorporation of LeuD3. Incorporation was verified by tandem mass spectrometry (as described below) and LeuD3 incorporation efficiency (%) was calculated using the

formula = [H/L] median/(1 + [H/L] median)*100 (14). An incorporation efficiency of > 98% of LeuD3 was achieved within five weeks.

## Acid stress exposure of *M. tuberculosis* for phosphoproteome and proteome analysis.

Experiments were performed during mid-exponential phase ($OD_{600nm}$ of 0.8-1.0). SAMMtb cultured in LeuD3 was exposed to pre-warmed modified Sauton's broth without Tween-80 at pH 4.5, whereas *M. tuberculosis* cultured in Leu0 served as control samples (i.e. pre-warmed Sauton's broth without Tween-80 at pH 7.0). A label swop was introduced to account for possible metabolic changes during label incorporation. SAMMtb cells were exposed to acidic (pH 4.5) and control (pH 7.0) conditions for 45 min and 24 hours. Experiments were performed in biological duplicate for phosphoproteomics analysis and biological triplicate for proteomics analysis. SAMMtb cells were pelleted (4 000 xg, 10 min, 4°C) at each time point and washed twice with ice-cold PBS pH 7.4. Cells were subsequently stored at -80°C until further use.



## Pulse-chase in *M. tuberculosis*.

Experiments were performed in biological duplicate, and during the mid-exponential phase of growth ($OD_{600nm}$ of 0.8-1.0). SAMMtb was cultured in modified Sauton's broth, containing 0.05% Tween-80, 24 µg/ml pantothenate and 50 µg/ml Leu0. Cells were centrifuged (4 000g, 10 min, 25˚C) and washed twice with pre-warmed PBS pH 7.4 before exposure to pre-warmed modified Sauton's broth, without Tween-80 but containing 50 µg/ml LeuD3, at pH 7.0 and 4.5, respectively. Cultures were harvested by centrifugation (4 000g, 10 min, 4°C) at 0, 3, 6, 12, 24 and 48 hours. SAMMtb cells were washed twice with ice-cold PBS pH 7.4 and stored at -80°C until further use.

## Filter-aided sample preparation for phosphoproteome analysis.

Samples were processed using a modified version of the filter-aided sample preparation (FASP) method (15). Briefly, samples from each matched SILAC condition (i.e. control and acid stress) were mixed in a 1:1 ratio, and 2.6 mg total protein (1.3 mg from each SILAC state) was processed for phosphoproteome analysis. Combined samples were reduced with 5 mM tris-2-carboxyethyl-phosphine (TCEP) at 25°C for 1 hour, and subsequently alkylated with 5.5 mM iodoacetamide at 25°C for 1 hour in the dark. The reduced and alkylated samples were transferred to a 15 ml Amicon 30kDa cut-off spin filter (Millipore) and centrifuged (4 000 xg, 30 min, 25˚C). Two ml UA buffer (8 M urea in 50 mM TEAB) was added to the filter and the samples were centrifuged (4000 g, 30 min, 25˚C). This process was repeated twice for a total of three UA washes. Two ml 50 mM TEAB was added to the filter and the samples were centrifuged (4000g, 30 min, 25˚C). This process was repeated twice for a total of three TEAB washes. Proteins were

on-filter digested for 24 h at 37 °C, using a 1:50 ratio of trypsin (Promega) to protein. The filter unit was placed in a new collection tube after overnight digestion, and the peptides were obtained in the flow-through by centrifugation (4000 xg, 30 min, 25 °C). Peptides were concentrated by lyophilisation.

## Phosphopeptide enrichment.

Peptides were desalted using Sep-Pak C18 cartridges (Waters) and dried prior to phosphopeptide enrichment. Phosphopeptides were serially enriched from 2.5 mg starting material and the MagReSyn® Ti-IMAC microspheres (ReSyn Biosciences), the MagReSyn® $TiO_2$ microspheres (ReSyn Biosciences) and the MagReSyn® $ZrO_2$ microspheres (ReSyn Biosciences), respectively. Enrichments were performed as recommended by the manufacturer. Briefly, microspheres were resuspended by vortex mixing and placed in a magnetic separator. Microspheres were washed with 70% ethanol with gentle agitation for 5 minutes. This wash step was repeated once more. The microparticles were subsequently washed with 1% $NH_4OH$ and gentle agitation for 10 minutes. Microparticles were equilibrated with loading buffer (1M glycolic acid in 80% ACN, 5% Trifluoroacetic acid (TFA)), for 1 minute. This was followed by a total of three washes with loading buffer. For phosphopeptide enrichment, the equilibrated $TiO_2$ microparticles were resuspended in peptide samples dissolved in loading buffer. The samples were incubated for 20 minutes at room temperature with continuous mixing. Samples were placed in a magnetic separator, the supernatant was removed and sequentially added to pre-equilibrated $ZrO_2$ and Ti-IMAC microparticles, respectively. Samples were incubated as described above. Unbound peptides were removed by washing with loading buffer and gentle agitation for 1 minute. Non-specifically bound peptides were removed by washing with wash buffer (80% ACN and 1% TFA), for 2 minutes with gentle agitation. This step was repeated for a total of two washes. Two additional wash steps were performed using of 10% ACN and 0.2% TFA. The bound phosphopeptides were extracted using elution buffer (1% $NH_4OH$) and incubated for 15 minutes with continuous mixing. Three elution steps were performed. Samples were desalted using STAGE tips, prepared with Empore C18 SPE Disks (Sigma-Aldrich, St Louis, USA), and dried before storage at -20°C.

## Filter-aided sample preparation for proteome analysis.

One hundred micrograms of total protein (i.e. 50 µg per SILAC state) was processed for proteome analysis. Combined samples were reduced and alkylated, as described above. The reduced and alkylated samples were transferred to a 0.5 ml Amicon 30 kDa cut-off spin filter (Millipore), and centrifuged (14 000 xg, 15 min, 25 ˚C). Four hundred microliters UA buffer (8 M urea in 100 mM Tris-HCl, pH 8.5) was added to the filter and the samples were centrifuged (14 000g, 15 min, 25 ˚C). This process was repeated

twice for a total of three UA washes. Four hundred microlitres 50 mM TEAB was added to the filter and the samples were centrifuged (14 000g, 15 min, 25˚C). This process was repeated twice for a total of three TEAB washes. Proteins were on-filter digested for 24 h at 37°C, using a 1:50 ratio of trypsin to protein. The filter unit was placed in a new collection tube after overnight digestion and the peptides were obtained in the flow-through by centrifugation (14 000g, 15 min, 25˚C). The peptides were dried in a vacuum concentrator and stored at -80°C until further analysis.



## Peptide desalting and fractionation.

Peptides (40 µg) from the 45 min and 24 h acid exposure experiments, as well as the 0 h time point of the protein turnover experiment, were fractionated using STAGE-tip-based fractionation. Peptides were sequentially eluted from STAGE-tips using an increasing gradient of ACN in 10 mM TEAB (5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 25% and 50%). Non-consecutive fractions were concatenated to obtain 4 fractions. Fractions were then vacuum dried and stored at -80°C.

## LC-MS/MS analysis.

Peptides were dissolved in 2% acetonitrile, containing 0.1% formic acid, and each sample was analysed, independently, on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific), connected to a Thermo Scientific UltiMate 3000 RSLCnano System (Thermo Fisher Scientific). Peptide separation was performed on a PepMap 300 C18 LC pre-column (300 µm ID x 5 mm, 5 µm, 300 Å), followed by separation on an analytical column (75 µm ID x 350 mm) packed with C18 Aeris Peptide 3.6 µm beads (Phenomenex 04A-4507), at a flow rate of 300 nL/min. Solvent A was 2% acetonitrile in 0.1% formic acid, and solvent B was 100% acetonitrile in 0.1% formic acid. The gradient used was as follows: solvent B was maintained at 2% for 5 min, followed by an increase from 2 to 10% B in 5 min, 10-35% B in 65 min, 35-50% B in 10 min, 50-80% B in 0.1 min, and maintained at 80% B for 10 min. The Orbitrap Fusion was operated in positive ion data-dependent mode for Orbitrap-MS and Orbitrap-MS2 data acquisition. Data were acquired using the Xcalibur software package. The precursor ion scan (full scan) was performed in the Orbitrap in the range of 350-1500 m/z with a nominal resolution of 120 000 at 200 m/z. The number of accumulated ions was set to $4 \times 10^5$. Ion filtering for Orbitrap-MS2 data acquisition was performed using the quadrupole with a transmission window of 1.5 m/z. The most intense ions above an intensity threshold of $5 \times 10^3$ were selected for high-energy collisional dissociation (HCD). For proteome samples, an HCD normalized collision energy of 32.5% was applied to the most intense ions, and fragment ions were analysed in the Orbitrap at a resolution of 30 000. The settings for phosphoproteome analysis were the same, except that an HCD normalized collision energy of 28% was used. The number of Orbitrap-MS2 events between full

scans was determined on-the-fly to maintain a 3s fixed duty cycle. Dynamic exclusion of ions within a ±10 ppm m/z window was implemented using a 30 s exclusion duration. An electrospray voltage of 1.9 kV and capillary temperature of 280°C, with no sheath and auxiliary gas flow, was used. The automatic gain control (AGC) settings were $4 \times 10^5$ ions with a maximum ion accumulation time of 50 ms for Oribtrap-MS, and $5 \times 10^4$ ions with a maximum ion accumulation time of 50 ms for Orbitrap-MS2 scans, respectively. Ions with <2+ or undetermined charge state were excluded from MS2 selection

## Mass spectrometry data analysis.

All tandem mass spectra were analysed using MaxQuant 1.6.0.16 (16) and searched against a customized *M. tuberculosis* proteome database using the built-in Andromeda search engine (17,18). Custom database construction was performed as previously described (12). Briefly, proteins containing single amino acid variants, derived from WGS analysis, were concatenated to the *M. tuberculosis* reference proteome (downloaded from Uniprot on 26 September 2017, containing 3993 entries). The custom *M. tuberculosis* database contained 4012 entries. Peak list generation was performed within MaxQuant and searches were performed using default parameters. Strict criteria for peptide and protein identification was implemented to control the false discovery rate (FDR) at < 0.01. LeuD3 was specified as heavy label during SILAC data analysis. Carbamidomethylation of cysteine was set as fixed modification, and oxidized methionine and N-terminal acetylation was used as variable modifications in proteome searches. The same settings were used for phosphoproteome analysis with the addition of phosphorylation on serine, threonine, and tyrosine as variable modification. Two missed tryptic cleavages were allowed. The “match between runs” and “re-quantify” options were enabled during database searches. For the incorporation check and protein turnover analysis, both these settings were ommited. Proteins were considered confidently identified when they contained at least 2 unique tryptic peptides per protein. Proteins that contained similar peptides, that could not be differentiated based on tandem mass spectrometry analysis alone, were grouped to satisfy the principles of parsimony. All contaminants, reverse hits, and proteins only identified by site were removed before downstream data analysis. Phosphorylation sites in phosphorylated peptides were considered as high confidence sites when they had a posterior error probability (PEP) of <0.01 and a localization probability of >0.75.

## Protein turnover data analysis.

SILAC ratios (non-normalized) obtained from MaxQuant were used for protein turnover analysis. The rate of light signal degradation and half-lives were calculated as previously described (19). Briefly, ratios were filtered for each protein to contain at least

three ratios across all five time points analysed, and were subsequently transformed using equation 1 to yield r, where r is the transformed ration and R is the raw ratio.

$$r = \ln(R + 1) \quad (1)$$

The coefficient of determination ($R^2$) was calculated after transformation of the data set and only values greater than 0.85 were used for further calculations. Next, we determined the slope (Kdeg) and corrected for signal dilution (Kdp) as a result of mycobacterial doubling, which is represented in equation 2 and 3, where $t_i$ represents each time point and *tcc* represents the time in hours for cell division to occur.

$$K_{deg} = \frac{\Sigma \ln(r) t_i}{\Sigma t_i^2} \quad (2)$$

$$K_{dp} = K_{deg} - \frac{\ln(2)}{tcc} \quad (3)$$

Using the slope, it is possible to derive a first order reaction equation and calculate half-life ($T_{1/2}$) as shown in equation 4.

$$T_{1/2} = \frac{\ln(2)}{K_{dp}} \quad (4)$$

From the degradation rate and half-life, certain classes were defined as previously described (20). Class I was defined as (Kdeg ≥ 2xKdill), class II as (0.5xKdill ≤ Kdeg ≥ 2xKdill) and Class III as (Kdeg ≤ 2xKdill). This translates to proteins that have a degradation rate twice as fast as the dilution rate (Class I), proteins that degrade faster than half the mycobacterial doubling time (Class III) and proteins that fall between these two definitions (Class II).

## RESULTS

### Time-dependent proteome analysis of *M. tuberculosis* during acid stress

Acidification of the phagolysosome by macrophages is an important early strategy to kill and process phagocytosed pathogens for further immune response (21). The lack of macrophage acidification in *M. tuberculosis* can be tied to the absence of a vacuolar proton-ATPase (22,23). For continued growth, *M. tuberculosis* maintains the intracellular pH at an approximate pH of 6.2 and a pH below 5 can arrest growth (24,25). Activation of the immune system by interferon gamma can overcome the acidification block imposed by *M. tuberculosis* and acidify the compartment to a pH range between 4.5 and 5.0. In *in vitro* experiments the intracellular pH of *M. tuberculosis* is maintained at

neutral even in the presence of highly acidic environments (26). This indicates that *M. tuberculosis* is able to survive the pH drop and is also metabolically active and thus able to maintain the intracellular pH. Furthermore, studies have shown that *M. tuberculosis* can maintain this state for extended periods of time and rapidly restart transcription once optimal conditions are restored (27–29).

We sought to identify downstream proteome changes in SAMMtb during acid stress. Heavy labelled mycobacteria were exposed to acidic conditions (pH 4.5) or physiological pH (pH 7.0) for 45 min and 24 h. The proteomes were harvested and the contents were analysed by tandem mass spectrometry. Principal component analysis revealed a clear separation in the first component between the two time points and little to no separation in the second component (Figure 1A). This indicates that there is little variation between replicates and the majority of variation is driven by the duration of the experiment. To find the source of this variation, we performed hypothesis testing on the 2093 protein groups, and found significant upregulation of 23 proteins while 322 were significantly downregulated at 24 hours exposure (Figure 1B, Data S1, Table 1). As there is a large number of downregulated proteins, we used gene ontology enrichment analysis to extract the major processes involved using a hypergeometric test. To identify a common role for the downregulated proteins, we only considered the child terms from the biological processes. A total of 172 nodes from biological processes (2173 total nodes) were identified where the top 10 most enriched terms fell within this parent node all with an odds ratio above 1 and a mean odds ratio of 10 (Supplemental figure 1). These were mainly associated with DNA and processes surrounding its synthesis and degradation. DNA is susceptible to acidic stress and these processes are likely in place to prevent DNA damage from accumulating (30).

To gain more insight into this data, we extracted the top 20 most significantly up- and-downregulated proteins as identified after 24 hour exposure and visualised the abundance change over time (Figure 1C). As predicted by the hypothesis testing, there was no difference in the protein expression of these proteins at 45 minutes of exposure. Interestingly, the majority of downregulated proteins formed part of the Dos regulon (31). Prominent members included Rv2623 and HspX which are both associated with the negative regulation of growth (32,33). In addition, a distinct cluster of upregulated proteins formed within the significantly regulated proteins (Figure 1C). These included TcrX (a member of the two component regulatory system), Rv2558 (involved in the cellular response to starvation), PrpD (catabolism of fatty acid via TCA cycle) and Rv3174 (a short chain dehydrogenase). Furthermore, Rv1425 (involved in the stress response to hypoxia and nitric oxide), and MymA (involved in to maintaining cell membrane permeability and mycolic acid composition during acid stress) was present as well. The

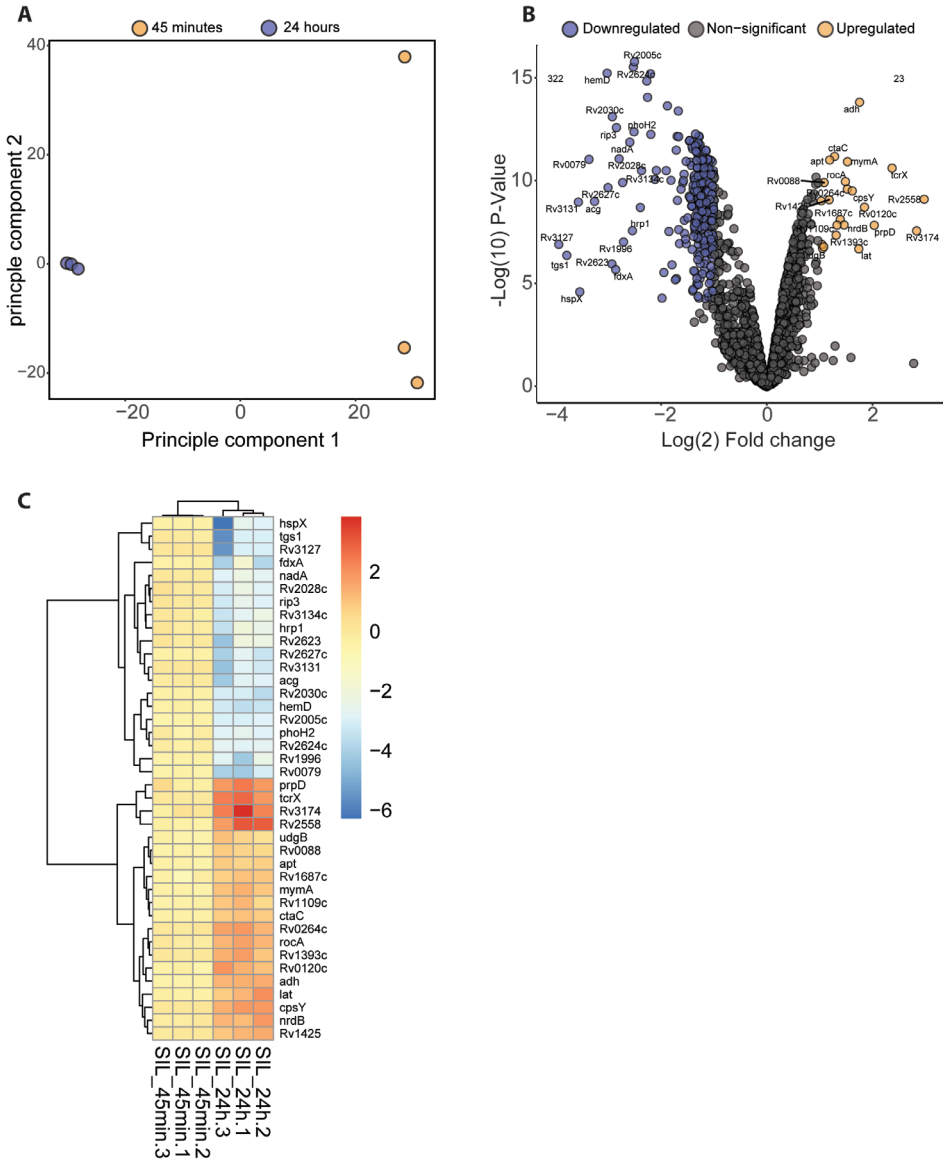



**Figure 1: Quantitative proteomics reveals major down regulation of proteins after 24 hours exposure pH 4.5. A)** Principle component analysis of SILAC ratio's at 45 minutes and 24 hours. **B)** Volcano plot depicting the fold change and corrected p-value of all proteins at 24 hours post stress compared to 45 minutes post stress. Significance was determined by hypothesis testing using Limma and Benjamini-Hochberg FDR was applied to the p-values to yield q-values. Points are coloured based on this q-value where a q-value below 0.05 as well as a absolute log fold change greater than 1 was considered significant. **C)** Heatmap depicting the top 20 up- and downregulated proteins by q-value and fold change. Data used for the generation of these graphs were obtained from three independent experiments.

two component system is responsible for cellular signalling by phosphorylation as well as mediating cascades in response to environmental signals (34,35). The increase of fatty acid metabolism is likely to account for the removal of toxic propionyl co-enzyme A, which accumulates during the catabolism of cholesterol (36,37). Fatty acid metabolism and cholesterol has been linked to *M. tuberculosis* carbon utilisation when grown in acidic conditions (38).

Taken together our results show that the bacilli is responding to the low pH by increasing production of virulence factors, utilising alternative energy stores and mediating signalling cascades. We also see a global down regulation occurring, which is in line with other observations where the bacilli slows down metabolism in preparation to enter a dormant state (29). This trend is a general one, where the majority of the proteins are downregulated, a process mediated by the Dos regulon (39–41). Interestingly, we found also found upregulated and thus active over time, although fewer in number compared to the downregulated profile. As *M. tuberculosis* is first and foremost an intracellular pathogen, it is thus unsurprising that the proteins produced in response to acidification are those needed to increase pathogenicity.

## Early exposure to acid stress remodels the *M. tuberculosis* phosphoproteome without affecting the proteome.

Phosphorylation plays an important role in *M. tuberculosis* as evident by 11 protein kinases and 12 two component systems (42,43). Phosphorylation cascades are used to activate or deactivate proteins (44). This can result in any number of changes and has been demonstrated to affect processes such as gene expression in *M. tuberculosis* through PknB (45). To identify the protein phosphorylation patterns, we exposed SAMMtb to acid stress (pH 4.5) and control (pH 7.0) for 45 minutes.

In total, we identified 512 phosphorylated peptides in SAMMtb during exposure to acid stress and control conditions. Two hundred and eighty-nine phosphorylation sites had a posterior error probability (PEP) score of <0.01 and a localization probability of >0.75, which were deemed to be high confidence identifications (46). These phosphorylation sites were mainly localised to serine and threonine residues (Supplementary figure 2A). By comparing the protein abundance to the phosphorylation abundance we found that the majority coincide, indicating that the phosphorylation is proportionate to the protein content (Figure 2A). Thirty-five phosphorylation sites were regulated by ≥ 2-fold change after normalising phosphorylation sites ratios to protein ratios (Figure 2B, supplementary figure 2 B1 – B35). The 35 phosphorylation sites mapped to 33 phosphoproteins that are involved in diverse biological processes (Data SI, Table 2).

Twenty out of the 35 regulated phosphorylation sites we identified, have previously been reported in mycobacteria (47–49). The protein with the most abundant phosphorylation relative to protein abundance was Rv3407 (VapB47). This protein is part of the toxin-antitoxin (TA) protein family and is expressed to counter the effects of the VapC47 toxin (50). The TA systems is closely linked to persisters and non-replicating states in bacteria (51–54). Previous studies have demonstrated that expression of a VapC *M. tuberculosis* homolog results in a lack of translation which is likely due to RNAse activity (50,55). As we do not detect VapC47 within our phosphorylation proteomics data it is likely that the increased phosphorylation is stabilising the anti-toxin which in turns neutralises the toxin counterpart. Indeed we do see that the majority of VapC proteins detected at 24 hours are downregulated compared to their counterparts at 45 minutes (Figure 1B, Data S1, table 1). As acidic stress can induce DNA damage, this enzyme is likely detoxifying nucleotides to remain viable within the environment. Due to the low number of regulated phosphorylation sites, typical hypergeometric statistics would not identify any significantly enriched groups. As an alternative, we performed functional annotation using the definitions provided by Mycobrowser. From this we could deduce that a large proportion of the phosphoproteins were localised to the cell surface and are involved in protein secretion as well as translocation across the membrane (EccCb1, EccC3, EsxB, Rv0394c, TatB, Mmps3). Other phosphoproteins, with regulated phosphorylation sites, are associated with roles in signal transduction and transcriptional responses (SigH, PknD, RpsA, and GreA), carbohydrate metabolism (GlcB), cell wall and lipid metabolism (GlmM, Pks13, CmaA2), cellular homeostasis (MoeA1, Rv0688, TrxB),and regulation of cell growth (VapB46, VapB47) (figure 2C). While this is an early timepoint, the increased phosphorylation of target proteins likely indicate a cellular response to environmental change.

Here we find that increased phosphorylation is primarily associated with the membrane itself and regulated by protein kinases. In our results we find PknD has phosphorylation sites with more than two fold regulation and PknH has phosphorylation sites between 1.5 and two fold. Both PknH and PknD have transmembrane sensory domains which transduce extracellular signals across the membrane to the cytoplasm (42,56). There is indication that PknH is involved in the negative regulation of growth, where deletion of this protein results in higher bacillary load *in vivo* (57), a phenotype we observe indirectly at 24 hours. We show widespread changes in the phosphoproteome of *M. tuberculosis* during early exposure to acid stress, while the proteome is yet to change during this initial exposure. Phosphorylation changes on S/T/Y residues, together with other signalling cascades such as histidine phosphorylation of two-component regulatory systems, will induce survival mechanisms in response to changes in the environment.

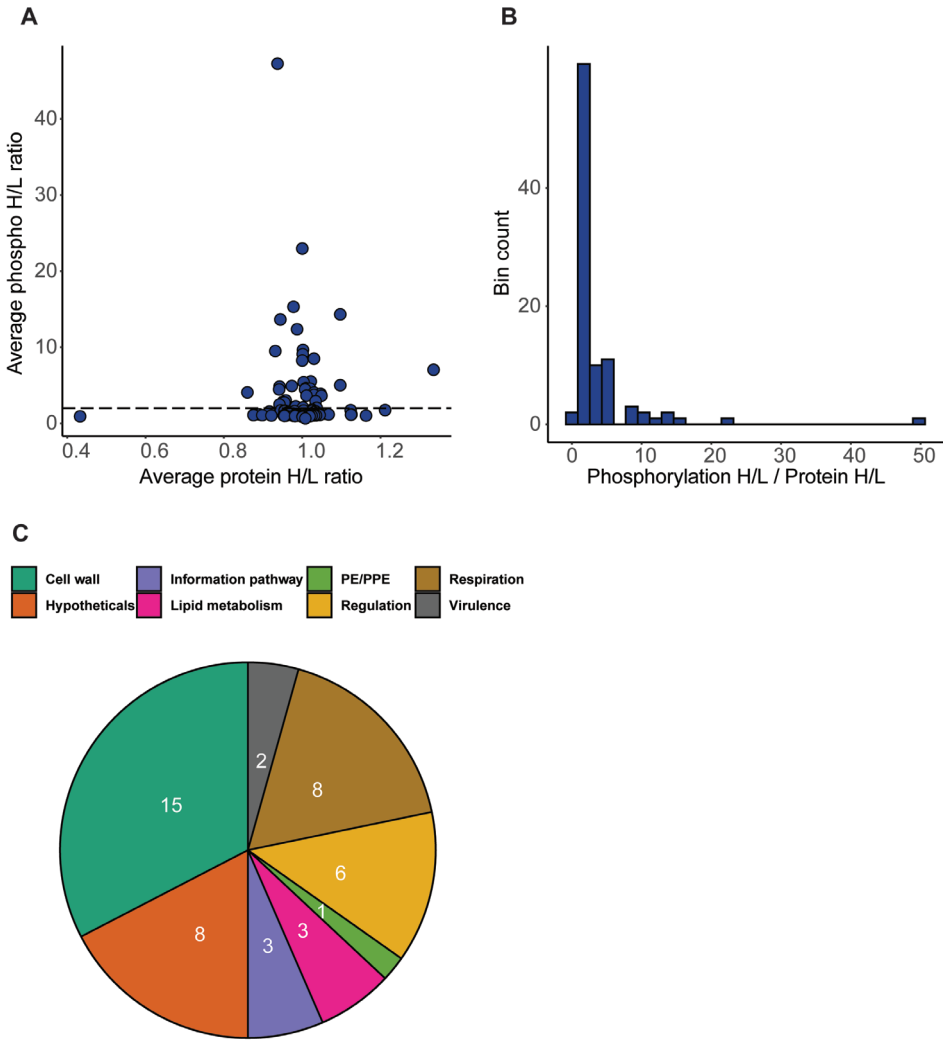



**Figure 2: Phosphorylation is dynamic and can be upregulated regardless of protein abundance in response to environmental stress. A)** Scatter plot depicting the phosphorylation abundance relative to the protein abundance of the phosphorylated peptides at 45 minutes post stress. The horizontal line indicates the fold change cut off which was set to 2. The abundance of phosphorylated proteins is greater than their protein abundance to which they belong. This indicates increased phosphorylation relative to the protein abundance. **B)** Bar graph indicating the distribution of phosphorylated peptide abundance in relation to protein abundance. The greater this ratio the more phosphorylation is found on the phosphorylated peptides relative to their proteins to which they belong. **C)** Pie chart indicating the number of proteins with regulated phosphorylation sites at 45 minutes post exposure and their corresponding Mycobrowser annotation. The majority of these proteins were associated with the mycobacterial cell wall. The data used to generate these graphs were obtained from two independent experiments.

## *M. tuberculosis* pulse-chase SILAC reveals steady state turnover of type VII secretion proteins and shutdown of amino acid transport upon acid stress.

Protein turnover is a balance between protein synthesis and degradation, which is responsible for maintaining protein homeostasis (proteostasis). Proteostasis is an important, and often overlooked, cellular function that maintains proteome integrity by recycling old or damaged proteins. Proteostasis can be shifted to accommodate cellular needs in response to environmental stimuli. We used pulse SILAC labelling to investigate the rate of global protein turnover, i.e. measurement of newly synthesized proteins by tracking Leu0 signal decay over five distinct time points.

During growth at pH 7.0, leucine was increasingly incorporated, with 183 protein groups containing heavy leucine after 3h of incubation and 800 protein groups containing heavy leucine after 48h of incubation (Figure 3A). We could identify three distinct classes of proteins based on protein turnover during growth at pH 7.0. These classes were was first defined by a study investigating protein turnover in *S. cerevisiae* and *S. pombe* during growth at physiological pH (20). The three classes represent the rate of turnover as a function of protein dilution that occurs naturally due to growth. Briefly, class I represents short-lived proteins (half-lives less than 13h), class II represents proteins with intermediate half-lives (half-lives between 13 and 57 hours), and class III represents long-lived proteins (greater than 57 hours or incalculable). These classes are defined based on their half-life distribution which is calculated from the degradation rate. At pH 7.0, most proteins were associated with a low corrected degradation rate (Kdp) (class III turnover rate), while the least amount of proteins had a fast degradation rate (class I) (Figure 3B). In total, six proteins were identified with a rapid turnover, 42 proteins had an intermediate turnover and 478 proteins were long lived at pH 7.0 (Figure 3C). The proteins with rapid turnover were involved in lipid metabolism (Ino1 and PapA1), fatty acid metabolism (AccD6), transcription (Rv1830), DNA replication (DnaE1) cell growth (VapB47) (Data SI, Table 3). This observation was unsurprising as the bacilli are actively growing and processes such as transcription and replication are of importance to cellular survival (58). *M. tuberculosis* DnaE1 is responsible for the proofreading activity and dysfunction of this protein drastically increases the mutation rate (59). The rapid turnover of this protein is likely required to maintain its fidelity and proofreading capabilities (59). Proteins that provide anti-toxin properties also require high maintenance as dysfunctional anti-toxin components can result in self toxicity. The class II and III turnover groups had many associated proteins and we therefore performed gene ontology enrichment analysis. From the class II enrichments, 3 of the top 10 most enriched terms were associated with protein secretion and 2 of these terms were associated with interspecies interactions. (Supplementary figure 3A). The class II

enrichments are indicative of proteins and the processes they collectively represent in steady state. This data thus demonstrates an active growth cycle for the type VII secretion system, where the components or effectors are actively maintained. As these proteins are membrane bound and the bacteria is not under stress that these proteins would undergo a slow turnover. However, *M. tuberculosis* is a specialized pathogen and a steady state of these proteins may indicate their importance as first response elements to outside stressors. There is a noticeable delay between transcription and translation in *M. tuberculosis,* it is thus important that the proteins associated with combatting host defences are not delayed in their synthesis (60). For the long-lived class III proteins, all the top 10 most enriched terms were associated with metabolism and growth of the organism, with growth and nitrogen metabolism representing the most enriched terms (Supplementary figure 3B). The long lived proteins can display a negative half-life, due to the measured H/L ratio being smaller than expected for proteins that are not degraded within the first doubling. It is certain that the proteins which have negative half-lives are the most stable. While metabolism was highly enriched it was not informative as these processes were too general to draw conclusions from.

Surprisingly, there was very little incorporation of LeuD3 in the proteome of SAMMtb during exposure to pH 4.5 over 48 hours. A total of 19 protein groups contained signal indicative of heavy leucine incorporation (Figure 3D). From these, we could calculate the half-life of two proteins (Data SI, Table 4). The minimal leucine uptake under acidic stress indicates that SAMMtb is unable to efficiently import leucine from the growth media. Whether this inability is due to SAMMtb becoming metabolically inactive or caused by deprotonation and therefore affecting amino acid transporters remains to be determined. Therefore, protein synthesis that occurs during acidic stress in SAMMtb must rely on amino acids imported early during exposure to acid stress or before. The integrity of the proteome must thus be maintained by salvaging amino acids from other proteins or by biosynthesis only. As leucine biosynthesis is blocked in the SAMMtb strain the source of leucine to sustain expression is likely derived from the remaining free leucine available within the cell or through biosynthesis from another source. Previous studies by *Gouzy et al* demonstrated a role for asparagine transport during acidic stress. In this scenario asparaginase is released to the extracellular milieu which hydrolyses asparagine to aspartate and ammonia. The ammonia buffers the extracellular environment and aspartate can be imported by AnsP transporter (61). The imported aspartate can follow metabolic pathways to synthesise other amino acids and provide free nitrogen to the pool (62). It is likely that SAMMtb under stress follows a similar pathway and favouring aspartate uptake rather than leucine uptake.

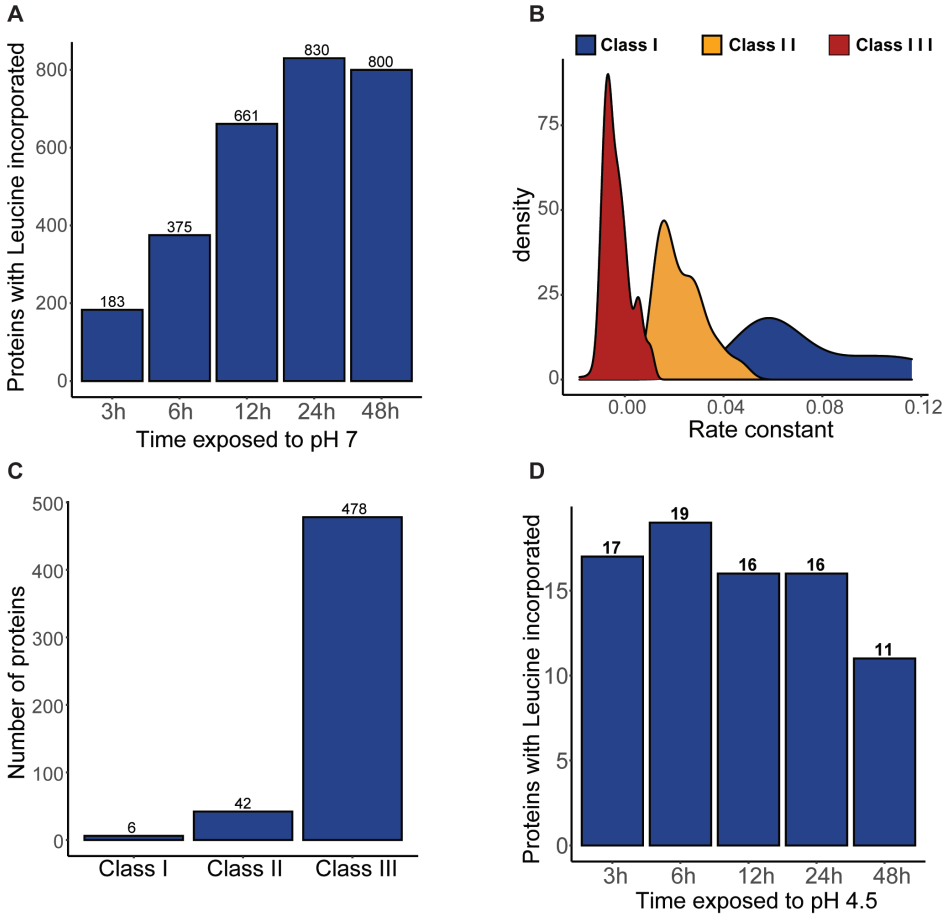



**Figure 3: Pulse chase SILAC reveals a shutdown of leucine import upon exposure to acid stress.** **A)** Incorporation of leucine at neutral pH (pH 7) at 3 hours, 6 hours, 12 hours, 24 hours and 48 hours. Leucine is incorporated steadily within one cell cycle indicating amino acid take up from the environment and active protein synthesis. Values atop the bar graph indicate the number of proteins with incorporated leucine. **B)** Density plot indicating the rate constant of each protein as well as the classes to which the various proteins belong. The rate constant is used to calculate the half-life of each protein. Class I was defined as (degradation rate ≥ 2 x dilution rate), class II as (0.5 x dilution rate ≤ degradation rate ≥ 2 x dilution rate) and Class III as (degradation rate ≤ 2 x dilution rate) where dilution rate was set to 28 hours. **C)** Bar graph depicting the number of proteins that correspond to each protein turnover class where the values are indicative of the protein count. The majority of proteins are associated with a slow protein turnover during active growth. **D)** Incorporation of leucine at low pH (pH 4.5) at 3 hours, 6 hours, 12 hours, 24 hours and 48 hours. A maximum of 19 proteins had heavy leucine incorporated across all time points and thus protein turnover could not be calculated for bacteria exposed to stress. Ratio's were calculated from two independent experiments, in the cases where one of the replicate had a missing value the corresponding value was used.

The observations made in the protein turnover of SAMMtb provided important insights into the behaviour of *M. tuberculosis* during active growth as well as acidic stress. The most striking observation was almost immediate shutdown of leucine transport during acid stress. This has interesting implications for dormancy and how the cell functions to sustain protein synthesis. More direct observations in protein turnover were made at pH 7. At this pH, a notable result was the use of the type VII protein secretion systems which were in steady state during active growth. Thus this secretion system and interspecies interaction forms an important process for *M. tuberculosis* regardless of the level of stress. Finally the long lived proteins were responsible for cellular metabolism during optimal growth. This is also to be expected as the processes required to maintain sustained growth and accumulation of biomass (20), which is the staple of active growth.

## DISCUSSION

Here we have performed an exhaustive investigation into *M. tuberculosis* using metabolic labelling on a leucine and pantothenate auxotroph of *M. tuberculosis* (12,13). Auxotrophy allows for metabolic labelling of the total proteome using heavy isotopes of leucine from which various quantitative proteomics studies are possible. Here we were able to exploit this technique to determine the temporal proteome effects, dynamic quantitative phosphorylation sites as well as protein homeostasis. Acidification of the phagosome is an important innate immune process and necessary for bacterial clearance. The enzymes involved in lysosomal degradation as well as production of reactive oxygen species and nitrogen species are optimal at low pH (63,64). *M. tuberculosis* is a facultative intracellular pathogen and has evolved to circumvent host defences such as acidification of the phagolysosome (21). Previous studies have also shown that blocking of the macrophage acidification is a prerequisite for ESX-1 dependent phagosomal escape (65). *M. tuberculosis* is also able to survive severe acidic conditions that is lethal to intracellular pathogens (66,67). Survival in low pH is clearly an important strategy employed in *M. tuberculosis* and a comprehensive overview of the various mechanisms has been reviewed previously (25). We investigated the mycobacterial response to low pH using the aforementioned techniques in order to understand how the pathogen reacts to this stress condition.

We have previously assessed the response of the auxotrophic *M. tuberculosis* strain to low pH and found upregulation of dormancy related proteins, indicating that the entrance into dormancy (12). Similar phenotypes have also been previously described and characterized by the upregulation of 48 genes (39–41,68). As the direct compari-

son between acidic pH and neutral pH has been extensively studied, we opted to use SILAC to rather study the temporal effects directly. We find a global decrease in the cellular response over time, with multiple members of the top 20 downregulated proteins associated with the dormancy regulon. These members included proteins responsible for mediating the negative regulation of mycobacterial growth such as Rv2623, Rv2626c and HspX (32,33). At first glance, it is curious that downregulation of multiple dormancy related proteins occurs at 24 hours compared to 45 minutes in our SILAC data. However, in our SILAC experiments we measure the collective effect of both proteomes over time. Our results thus show that while there are changes occurring due to the stress response, the temporal changes indicate a global decrease in protein abundance. The growth rate of *M. tuberculosis* is significantly retarded in response to acid stress and a global down regulation of proteins is thus an indication of metabolic shutdown, a known feature of the dormancy phenotype (69,70). Interestingly, we find dormancy related proteins upregulated at the 24 h mark as well. Previous studies have demonstrated a low abundance of stable transcripts while in a dormant state (29). While only within the first 24 hours, we too observe a cluster of upregulated proteins likely involved in maintaining the viability of the bacilli while transitioning into dormancy. The most prominent upregulated proteins included a member of the two-component system, TcrX. Deletion of the TcrX/TcrY system was found to result in increased bacillary growth implying their role in suppressing growth (71). As there is a global downregulation of genes, including others involved in suppressing growth, the TcrX/TcrY system may be involved in sustaining the signal to suppress growth. In addition, fatty acid metabolism and features of cholesterol metabolism were detected as well. Previous studies have shown that *M. tuberculosis* metabolizes cholesterol and other lipoproteins within macrophages (72). This response indicates an attempt by *M. tuberculosis* to utilize a specific carbon source during acidic conditions. As there is a global shutdown in progress and growth arrest occurs at the pH where these experiments are conducted, this may give indication that *M. tuberculosis* preferentially will utilize fatty acids when confronted with environments mimicking activated macrophages (24,25). Indeed, isocitrate lyase, an enzyme involved in metabolizing acetyl-CoA from fatty acids, is required for sustained infection (73).

We also detected multiple virulence factors albeit below the 2-fold change cut off. These included EccD2 and EccCa1 of the ESX secretion systems. The ESX-1 secretion system is well known for its role in combating acidification of host vacuoles (6,74–76). At 45 minutes, we found increased phosphorylation of EccCb1, the partner of EccCa1 (77) . The EccC component of the Type VII secretion systems is an ATPase and increased phosphorylation is a likely indicator of increased activity of the entire secretion system. EsxB allosterically binds to the ATPase pocket domains of EccC which

results in an increased ATPase activity (78). It is thus likely that the free phosphates are subsequently binding to both EccC and EsxB and are as a result detected at greater propensity within the phosphoproteome. As EsxA has a role in controlling phagosomal acidification, the increased activity of the ESX-1 system is expected upon sensing environmental pH change (65,79,80). Strikingly, this change is observed and implemented before any increases in the abundance of ESX components or effectors. It is tempting to speculate about the function that phosphorylation of ESX components may entail. Partial acidification is key for successful escape from the phagosome and phosphorylation may be the mechanism used to determine the optimal conditions for escape (81). Indeed there is a significant overlap in *M. tuberculosis* expression profiles *in vivo* and when subjected to pH fluctuation thereby indicating a substantial role for pH as an environmental signal (2). It is thus clear that this system plays an extremely important role in mediating the bacterial response to its environment. In addition, we found the protein turnover rate of type VII secretion components to be within steady state. This ensures that there is a steady supply of ESX components ready to increase or decrease in activity upon necessity by the bacilli and further demonstrates the importance of the type VII secretion system to the bacilli.

We were unable to calculate protein turnover rates in the acid stressed bacilli due to a lack of leucine uptake. As protein turnover was calculable at neutral pH, the acidic extracellular milieu must have an effect on leucine transport across the membrane. We previously speculated that acidic environments may impact leucine import, the reason behind the lack of leucine incorporation is however still unclear (12). In *E. coli*, leucine transport is mediated by a sodium symporter (82). Low pH results in an increase of protons in the extracellular milieu which may interfere with the amino acid import by abolishing a charge differential across the membrane. The mechanism of leucine transport is as of yet unknown in *M. tuberculosis.* It is therefore difficult to comment on the exact processes responsible for blocking leucine import in such a drastic manner. Previous studies have shown that aspartate is the favoured amino acid during infection and is produced by utilising an extracellular asparaginase to hydrolyse asparagine to aspartate and ammonia (61). It is likely that *M. tuberculosis* obtains nitrogen from this pathway which can be used to synthesize leucine instead of direct uptake from the environment and the ammonia to buffer the acidic pH. Whether *M. tuberculosis* simply does not import leucine during low pH compared to a more efficient system or is unable to do so remains unclear. Indeed, this model is reliant on the leucine biosynthesis pathway *in vivo* and not direct uptake (13). Another explanation to the lack of leucine incorporation may be the decreased protein synthesis as observed in Figure 1B. Leucine import may be unimpaired but the isotopes my not be incorporated into

the proteome. This is however less likely as we do observe differential regulation even at late time points when comparing low pH to neutral pH directly (12).

Here we have studied multidimensional proteome dynamics in response to acid stress, encompassing changes in temporal abundance, dynamic phosphorylation and protein turnover. We find that *M. tuberculosis* undergoes a large scale protein remodelling when exposed to acidic stress, which is driven in part by early phosphorylation cascades. We find the type VII secretion system to form a central role in early responses to an acidified environment and that this system is kept at a steady state during active growth. While we were unable to measure protein turnover comparatively to acidic stress, a shutdown of leucine transport was observed. This indicates a shift in major cellular processes when exposed to harsh environments. By investigating all aspects of the mycobacterial response to common stressors can we gain sufficient insight into the processes involved in infection. The SAMMtb strain certainly allows for such investigations, and has the potential to play a pivotal role in future mycobacterial research.

## ACKNOWLEDGEMENTS

JL would like to acknowledge DST/NRF Desmond Tutu Doctoral Program for financial assistance.

TH was supported by a South African National Research Foundation-Department of Science and Technology Innovation Postdoctoral Fellowship (SFP13071721852).

SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NRF.

The authors acknowledge the SA MRC Centre for TB Research and DST/NRF Centre of Excellence for Biomedical Tuberculosis Research for financial support for this work

## AUTHOR CONTRIBUTIONS.

**JLG**: methodology, formal analysis, investigation, data curation, writing

**PM**: formal Analysis, investigation, writing

**SLS**: conceptualisation, resources, writing, supervision, funding acquisition

**TH**: conceptualisation, methodology, formal analysis, investigation, writing, supervision, funding acquisition

## SUPPLEMENTARY DATA

Supplementary data for not in this document can be accessed at the following URL with a Mendeley account:

https://tinyurl.com/5u3sumvm

## REFERENCES


1. World Health Organisation. GLOBAL TUBERCULOSIS REPORT 2020 [Internet]. 2020 [cited 2020 Nov 24]. Available from: http://apps.who.int/bookorders.
2. Rohde KH, Abramovitch RB, Russell DG. Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues. Cell Host Microbe [Internet]. 2007 Nov 15 [cited 2020 Nov 24];2(5):352–64. Available from: https://pubmed.ncbi.nlm.nih.gov/18005756/
3. MacGurn JA, Cox JS. A genetic screen for Mycobacterium tuberculosis mutants defective for phagosome maturation arrest identifies components of the ESX-1 secretion system. Infect Immun [Internet]. 2007 Jun [cited 2020 Nov 24];75(6):2668–78. Available from: https://pubmed.ncbi.nlm.nih.gov/17353284/
4. de Chastellier C. The many niches and strategies used by pathogenic mycobacteria for survival within host macrophages. Immunobiology. 2009 Jul 1;214(7):526–42.
5. Ehrt S, Schnappinger D. Mycobacterial survival strategies in the phagosome: defence against host stresses. Cell Microbiol [Internet]. 2009 Aug [cited 2019 Jul 9];11(8):1170–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19438516
6. Russell DG. Mycobacterium tuberculosis: Here today, and here tomorrow [Internet]. Vol. 2, Nature Reviews Molecular Cell Biology. Nat Rev Mol Cell Biol; 2001 [cited 2020 Nov 24]. p. 569–77. Available from: https://pubmed.ncbi.nlm.nih.gov/11483990/
7. Russell DG. Mycobacterium tuberculosis and the intimate discourse of a chronic infection [Internet]. Vol. 240, Immunological Reviews. Immunol Rev; 2011 [cited 2020 Nov 24]. p. 252–68. Available from: https://pubmed.ncbi.nlm.nih.gov/21349098/
8. Schaible UE, Sturgill-Koszycki S, Schlesinger PH, Russell DG. Cytokine activation leads to acidification and increases maturation of Mycobacterium avium-containing phagosomes in murine macrophages. J Immunol [Internet]. 1998;160(3):1290-1296. Available from: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.jimmunol.org/cgi-bin/Retreiver.cgi/v160n3/1290/1290-abs-frame.html%5Cnhttp://www.jimmunol.org/cgi/reprint/160/3/1290.pdf
9. Via LE, Fratti R a, McFalone M, Pagan-Ramos E, Deretic D, Deretic V. Effects of cytokines on mycobacterial phagosome maturation. J Cell Sci. 1998;111 ( Pt 7:897–905.
10. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between Protein and mRNA Abundance in Yeast. Mol Cell Biol [Internet]. 1999 Mar 1 [cited 2020 Nov 24];19(3):1720–30. Available from: https://pubmed.ncbi.nlm.nih.gov/10022859/
11. Marguerat S, Schmidt A, Codlin S, Chen W, Aebersold R, Bähler J. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell [Internet]. 2012 Oct 26 [cited 2020 Nov 24];151(3):671–83. Available from: /pmc/articles/PMC3482660/?report=abstract
12. Mouton JM, Heunis T, Dippenaar A, Gallant JL, Kleynhans L, Sampson SL. Comprehensive Characterization of the Attenuated Double Auxotroph Mycobacterium tuberculosisΔleuDΔpanCD as an Alternative to H37Rv. Front Microbiol [Internet]. 2019 Aug 20 [cited 2020 Nov 24];10(AUG):1922. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2019.01922/full
13. Sampson SL, Dascher CC, Sambandamurthy VK, Russell RG, Jacobs WR, Bloom BR, et al. Protection Elicited by a Double Leucine and Pantothenate Auxotroph of Mycobac-

terium tuberculosis in Guinea Pigs. Infect Immun [Internet]. 2004 May [cited 2020 Nov 24];72(5):3031–7. Available from: /pmc/articles/PMC387862/?report=abstract

14. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics [Internet]. 2002 [cited 2020 Nov 24];1(5):376–86. Available from: https://pubmed.ncbi.nlm.nih.gov/12118079/
15. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods [Internet]. 2009 [cited 2020 Nov 24];6(5):359–62. Available from: https://pubmed.ncbi.nlm.nih.gov/19377485/
16. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
17. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V., Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805.
18. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics [Internet]. 2014 Sep [cited 2019 Oct 13];13(9):2513–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24942700
19. Schwanhüusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature [Internet]. 2011 May 19 [cited 2019 Mar 7];473(7347):337–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21593866
20. Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. Cell Rep [Internet]. 2014 Dec 11 [cited 2019 Mar 6];9(5):1959–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25466257
21. Ip WKE, Sokolovska A, Charriere GM, Boyer L, Dejardin S, Cappillino MP, et al. Phagocytosis and Phagosome Acidification Are Required for Pathogen Processing and MyD88-Dependent Responses to Staphylococcus aureus . J Immunol [Internet]. 2010 Jun 15 [cited 2020 Nov 24];184(12):7071–81. Available from: /pmc/articles/PMC2935932/?report=abstract
22. Pethe K, Swenson DL, Alonso S, Anderson J, Wang C, Russell DG. Isolation of Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation. Proc Natl Acad Sci United States Am [Internet]. 2004 Sep 14;101(37):13642–7. Available from: http://www.pnas.org/content/101/37/13642.abstract
23. Sturgill-Koszycki S, Schlesinger PH, Chakraborty P, Haddix PL, Collins HL, Fok AK, et al. Lack of acidification in Mycobacterium phagosomes produced by exclusion of the vesicular proton-ATPase. Science (80- ) [Internet]. 1994 Feb 4;263(5147):678–81. Available from: http://science.sciencemag.org/content/263/5147/678.abstract
24. MacMicking JD, Taylor GA, McKinney JD. Immune Control of Tuberculosis by IFN-γ-inducible LRG-47. Science (80- ) [Internet]. 2003 Oct 24 [cited 2020 Oct 30];302(5645):654–9. Available from: https://pubmed.ncbi.nlm.nih.gov/14576437/
25. Vandal OH, Nathan CF, Ehrt S. Acid resistance in Mycobacterium tuberculosis [Internet]. Vol. 191, Journal of Bacteriology. American Society for Microbiology (ASM); 2009 [cited 2020 Nov 24]. p. 4714–21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2715723/

26. Vandal OH, Pierini LM, Schnappinger D, Nathan CF, Ehrt S. A membrane protein preserves intrabacterial pH in intraphagosomal Mycobacterium tuberculosis. Nat Med [Internet]. 2008 Aug [cited 2020 Nov 24];14(8):849–54. Available from: https://pubmed.ncbi.nlm.nih.gov/18641659/
27. Salina EG, Grigorov AS, Bychenko OS, Skvortsova Y V., Mamedov IZ, Azhikina TL, et al. Resuscitation of Dormant "Non-culturable" Mycobacterium tuberculosis Is Characterized by Immediate Transcriptional Burst. Front Cell Infect Microbiol [Internet]. 2019 Jul 30 [cited 2020 Nov 24];9:272. Available from: https://www.frontiersin.org/article/10.3389/fcimb.2019.00272/full
28. Trutneva KA, Shleeva MO, Demina GR, Vostroknutova GN, Kaprelyans AS. One-Year Old Dormant, "Non-culturable" Mycobacterium tuberculosis Preserves Significantly Diverse Protein Profile. Front Cell Infect Microbiol [Internet]. 2020 Jan 31 [cited 2020 Nov 24];10:26. Available from: https://www.frontiersin.org/article/10.3389/fcimb.2020.00026/full
29. Ignatov D V., Salina EG, Fursov M V., Skvortsov TA, Azhikina TL, Kaprelyants AS. Dormant non-culturable Mycobacterium tuberculosis retains stable low-abundant mRNA. BMC Genomics [Internet]. 2015 Dec 16 [cited 2020 Nov 24];16(1):954. Available from: http://www.biomedcentral.com/1471-2164/16/954
30. Sorokin VA, Gladchenko GO, Valeev VA. DNA protonation at low ionic strength of solution. Die Makromol Chemie [Internet]. 1986 May 1 [cited 2020 Nov 24];187(5):1053–63. Available from: http://doi.wiley.com/10.1002/macp.1986.021870502
31. Park H-D, Guinn KM, Harrell MI, Liao R, Voskuil MI, Tompa M, et al. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis. Mol Microbiol [Internet]. 2003 Apr 9 [cited 2018 Feb 28];48(3):833–43. Available from: http://doi.wiley.com/10.1046/j.1365-2958.2003.03474.x
32. Drumm JE, Mi K, Bilder P, Sun M, Lim J, Bielefeldt-Ohmann H, et al. Mycobacterium tuberculosis Universal Stress Protein Rv2623 Regulates Bacillary Growth by ATP-Binding: Requirement for Establishing Chronic Persistent Infection. Ramakrishnan L, editor. PLoS Pathog [Internet]. 2009 May 29 [cited 2018 Feb 28];5(5):e1000460. Available from: http://dx.plos.org/10.1371/journal.ppat.1000460
33. Yuan Y, Crane DD, Barry CE. Stationary phase-associated protein expression in Mycobacterium tuberculosis: Function of the mycobacterial α-crystallin homolog. J Bacteriol [Internet]. 1996 [cited 2020 Nov 24];178(15):4484–92. Available from: https://pubmed.ncbi.nlm.nih.gov/8755875/
34. Cheung J, Hendrickson WA. Sensor domains of two-component regulatory systems. Vol. 13, Current Opinion in Microbiology. Elsevier Current Trends; 2010. p. 116–23.
35. Groisman EA. Feedback Control of Two-Component Regulatory Systems. Annu Rev Microbiol [Internet]. 2016 Sep 8 [cited 2020 Nov 24];70(1):103–24. Available from: http://www.annualreviews.org/doi/10.1146/annurev-micro-102215-095331
36. Eoh H, Rhee KY. Methylcitrate cycle defines the bactericidal essentiality of isocitrate lyase for survival of mycobacterium tuberculosis on fatty acids. Proc Natl Acad Sci U S A [Internet]. 2014 Apr 1 [cited 2020 Nov 24];111(13):4976–81. Available from: https://www.pnas.org/content/111/13/4976
37. Savvi S, Warner DF, Kana BD, McKinney JD, Mizrahi V, Dawes SS. Functional characterization of a vitamin B12-dependent methylmalonyl pathway in Mycobacterium tuberculosis: Implications for propionate metabolism during growth on fatty acids. J Bacteriol [Inter-

net]. 2008 Jun 1 [cited 2020 Nov 24];190(11):3886–95. Available from: http://jb.asm.org/. http://jb.asm.org/

38. Baker JJ, Johnson BK, Abramovitch RB. Slow growth of Mycobacterium tuberculosis at acidic pH is regulated by phoPR and host-associated carbon sources. Mol Microbiol [Internet]. 2014 Oct 1 [cited 2020 Nov 24];94(1):56–69. Available from: /pmc/articles/PMC4177513/?report=abstract
39. Boon C, Dick T. Mycobacterium bovis BCG response regulator essential for hypoxic dormancy. J Bacteriol [Internet]. 2002 Dec 15 [cited 2020 Nov 24];184(24):6760–7. Available from: http://jb.asm.org/
40. Leistikow RL, Morton RA, Bartek IL, Frimpong I, Wagner K, Voskuil MI. The Mycobacterium tuberculosis DosR regulon assists in metabolic homeostasis and enables rapid recovery from nonrespiring dormancy. J Bacteriol [Internet]. 2010 Mar [cited 2020 Nov 24];192(6):1662–70. Available from: https://pubmed.ncbi.nlm.nih.gov/20023019/
41. Oxygen, Nitric Oxide, and Carbon Monoxide Signaling [Internet]. [cited 2020 Nov 24]. Available from: https://www.caister.com/hsp/abstracts/mycobacterium/06.html
42. Prisic S, Husson RN. Mycobacterium tuberculosis Serine/Threonine Protein Kinases. Microbiol Spectr [Internet]. 2014 Oct 3 [cited 2020 Nov 24];2(5). Available from: /pmc/articles/PMC4242435/?report=abstract
43. Parish T. Two-Component Regulatory Systems of Mycobacteria. Microbiol Spectr [Internet]. 2014 Feb 7 [cited 2020 Nov 24];2(1). Available from: https://pubmed.ncbi.nlm.nih.gov/26082118/
44. Ardito F, Giuliani M, Perrone D, Troiano G, Muzio L Lo. The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy (Review) [Internet]. Vol. 40, International Journal of Molecular Medicine. Spandidos Publications; 2017 [cited 2020 Nov 24]. p. 271–80. Available from: /pmc/articles/PMC5500920/?report=abstract
45. Alqaseer K, Turapov O, Barthe P, Jagatia H, De Visch A, Roumestand C, et al. Protein kinase B controls Mycobacterium tuberculosis growth via phosphorylation of the transcriptional regulator Lsr2 at threonine 112. Mol Microbiol [Internet]. 2019 Dec 1 [cited 2020 Nov 24];112(6):1847–62. Available from: /pmc/articles/PMC6906086/?report=abstract
46. Collins MO, Wright JC, Jones M, Rayner JC, Choudhary JS. Confident and sensitive phosphoproteomics using combinations of collision induced dissociation and electron transfer dissociation. J Proteomics [Internet]. 2014 May 30 [cited 2020 Nov 24];103(100):1–14. Available from: /pmc/articles/PMC4047622/?report=abstract
47. Nakedi KC, Calder B, Banerjee M, Giddey A, Nel AJM, Garnett S, et al. Identification of novel physiological substrates of mycobacterium bovis BCG protein Kinase G (PknG) by label-free quantitative phosphoproteomics. Mol Cell Proteomics [Internet]. 2018 Jul 1 [cited 2020 Nov 24];17(7):1365–77. Available from: /pmc/articles/PMC6030727/?report=abstract
48. Fortuin S, Tomazella GG, Nagaraj N, Sampson SL, Gey van Pittius NC, Soares NC, et al. Phosphoproteomics analysis of a clinical Mycobacterium tuberculosis Beijing isolate: Expanding the mycobacterial phosphoproteome catalog. Front Microbiol [Internet]. 2015 [cited 2020 Nov 24];6(FEB). Available from: /pmc/articles/PMC4322841/?report=abstract
49. Verma R, Pinto SM, Patil AH, Advani J, Subba P, Kumar M, et al. Quantitative Proteomic and Phosphoproteomic Analysis of H37Ra and H37Rv Strains of Mycobacterium tuberculosis. J Proteome Res [Internet]. 2017 Apr 7 [cited 2020 Nov 24];16(4):1632–45. Available from: https://pubmed.ncbi.nlm.nih.gov/28241730/

50. Ramage HR, Connolly LE, Cox JS. Comprehensive Functional Analysis of Mycobacterium tuberculosis Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. Rosenberg SM, editor. PLoS Genet [Internet]. 2009 Dec 11 [cited 2021 Mar 1];5(12):e1000767. Available from: https://dx.plos.org/10.1371/journal.pgen.1000767
51. Norton JP, Mulvey MA. Toxin-Antitoxin Systems Are Important for Niche-Specific Colonization and Stress Resistance of Uropathogenic Escherichia coli. PLoS Pathog [Internet]. 2012 Oct [cited 2021 Mar 1];8(10). Available from: https://pubmed.ncbi.nlm.nih.gov/23055930/
52. Van Acker H, Sass A, Dhondt I, Nelis HJ, Coenye T. Involvement of toxin-antitoxin modules in Burkholderia cenocepacia biofilm persistence. Pathog Dis [Internet]. 2014 [cited 2021 Mar 1];71(3):326–35. Available from: https://pubmed.ncbi.nlm.nih.gov/24719230/
53. Cheverton AM, Gollan B, Przydacz M, Wong CT, Mylona A, Hare SA, et al. A Salmonella Toxin Promotes Persister Formation through Acetylation of tRNA. Mol Cell [Internet]. 2016 Jul 7 [cited 2021 Mar 1];63(1):86–96. Available from: https://pubmed.ncbi.nlm.nih.gov/27264868/
54. Helaine S, Cheverton AM, Watson KG, Faure LM, Matthews SA, Holden DW. Internalization of salmonella by macrophages induces formation of nonreplicating persisters. Science (80-) [Internet]. 2014 Jan 10 [cited 2021 Mar 1];343(6167):204–8. Available from: http://science.sciencemag.org/
55. Zhu L, Phadtare S, Nariya H, Ouyang M, Husson RN, Inouye M. The mRNA interferases, MazF-mt3 and MazF-mt7 from *Mycobacterium tuberculosis* target unique pentad sequences in single-stranded RNA. Mol Microbiol [Internet]. 2008 Aug 1 [cited 2021 Mar 1];69(3):559–69. Available from: http://doi.wiley.com/10.1111/j.1365-2958.2008.06284.x
56. Sharma K, Chandra H, Gupta PK, Pathak M, Narayan A, Meena LS, et al. PknH, a transmembrane Hank's type serine/threonine kinase from *Mycobacterium tuberculosis* is differentially expressed under stress conditions. FEMS Microbiol Lett [Internet]. 2004 Apr 1 [cited 2020 Nov 24];233(1):107–13. Available from: https://academic.oup.com/femsle/article-lookup/doi/10.1016/j.femsle.2004.01.045
57. Papavinasasundaram KG, Chan B, Chung J-H, Colston MJ, Davis EO, Av-Gay Y. Deletion of the Mycobacterium tuberculosis pknH gene confers a higher bacillary load during the chronic phase of infection in BALB/c mice. J Bacteriol. 2005 Aug;187(16):5751–60.
58. Salghetti SE. Functional overlap of sequences that activate transcription and signal ubiquitin-mediated proteolysis. Proc Natl Acad Sci [Internet]. 2000 Mar 28 [cited 2020 Nov 24];97(7):3118–23. Available from: /pmc/articles/PMC16202/?report=abstract
59. Rock JM, Lang UF, Chase MR, Ford CB, Gerrick ER, Gawande R, et al. DNA replication fidelity in Mycobacterium tuberculosis is mediated by an ancestral prokaryotic proofreader. Nat Genet [Internet]. 2015 May 27 [cited 2020 Nov 24];47(6):677–81. Available from: https://pubmed.ncbi.nlm.nih.gov/25894501/
60. Cortes T, Schubert OT, Banaei-Esfahani A, Collins BC, Aebersold R, Young DB. Delayed effects of transcriptional responses in Mycobacterium tuberculosis exposed to nitric oxide suggest other mechanisms involved in survival. Sci Rep. 2017 Dec 1;7(1).
61. Gouzy A, Bottai D, Levillain F, Dumas A, Chastellier C De, Wu T, et al. Mycobacterium tuberculosis Exploits Asparagine to Assimilate Nitrogen and Resist Acid Stress during Infection. PLoS Pathog [Internet]. 2014 Feb 20;10(2):e1003928. Available from: http://dx.doi.org/10.1371%2Fjournal.ppat.1003928

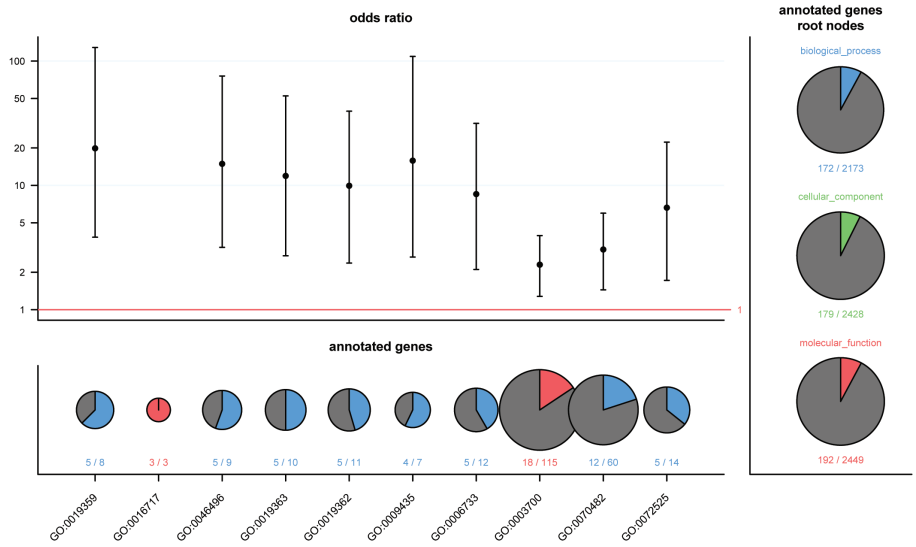
62. Gallant JLJL, Viljoen AJAJ, Van Helden PDPD, Wiid IJFIJF. Glutamate Dehydrogenase Is Required by Mycobacterium bovis BCG for Resistance to Cellular Stress. PLoS One. 2016;11(1):e0147706.
63. Sun-Wada GH, Tabata H, Kawamura N, Aoyama M, Wada Y. Direct recruitment of H+-ATPase from lysosomes for phagosomal acidification. J Cell Sci [Internet]. 2009 Jul 15 [cited 2020 Nov 24];122(14):2504–13. Available from: http://jcs.biologists.org/cgi/content/full/122/14/2504/DC1
64. Vieira O V., Botelho RJ, Grinstein S. Phagosome maturation: Aging gracefully [Internet]. Vol. 366, Biochemical Journal. Biochem J; 2002 [cited 2020 Nov 24]. p. 689–704. Available from: https://pubmed.ncbi.nlm.nih.gov/12061891/
65. Simeone R, Sayes F, Song O, Gröschel MI, Brodin P, Brosch R, et al. Cytosolic Access of Mycobacterium tuberculosis: Critical Impact of Phagosomal Acidification Control and Demonstration of Occurrence In Vivo. Salgame P, editor. PLOS Pathog [Internet]. 2015 Feb 6 [cited 2020 Nov 24];11(2):e1004650. Available from: https://dx.plos.org/10.1371/journal.ppat.1004650
66. Flannagan RS, Cosío G, Grinstein S. Antimicrobial mechanisms of phagocytes and bacterial evasion strategies [Internet]. Vol. 7, Nature Reviews Microbiology. Nat Rev Microbiol; 2009 [cited 2020 Nov 24]. p. 355–66. Available from: https://pubmed.ncbi.nlm.nih.gov/19369951/
67. Huynh KK, Grinstein S. Regulation of Vacuolar pH and Its Modulation by Some Microbial Species. Microbiol Mol Biol Rev [Internet]. 2007 Sep 5 [cited 2015 Oct 14];71(3):452–62. Available from: http://mmbr.asm.org/cgi/content/long/71/3/452
68. Mayuri, Bagchi G, Das TK, Tyagi JS. Molecular analysis of the dormancy response in Mycobacterium smegmatis : expression analysis of genes encoding the DevRâ€"DevS two-component system, Rv3134c and chaperone Î±-crystallin homologues . FEMS Microbiol Lett [Internet]. 2002 Jun [cited 2020 Nov 24];211(2):231–7. Available from: https://pubmed.ncbi.nlm.nih.gov/12076818/
69. Zhang Y. Persistent and dormant tubercle bacilli and latent tuberculosis [Internet]. Vol. 9, Frontiers in Bioscience. Frontiers in Bioscience; 2004 [cited 2020 Nov 24]. p. 1136–56. Available from: https://pubmed.ncbi.nlm.nih.gov/14977534/
70. Chao MC, Rubin EJ. Letting sleeping dos lie: Does dormancy play a role in tuberculosis? [Internet]. Vol. 64, Annual Review of Microbiology. Annu Rev Microbiol; 2010 [cited 2020 Nov 24]. p. 293–311. Available from: https://pubmed.ncbi.nlm.nih.gov/20825351/
71. Parish T, Smith DA, Kendall S, Casali N, Bancroft GJ, Stoker NG. Deletion of two-component regulatory systems increases the virulence of Mycobacterium tuberculosis. Infect Immun [Internet]. 2003 Mar 1 [cited 2020 Nov 24];71(3):1134–40. Available from: http://iai.asm.org/
72. Pandey AK, Sassetti CM. Mycobacterial persistence requires the utilization of host cholesterol. Proc Natl Acad Sci U S A [Internet]. 2008 Mar 18 [cited 2020 Nov 24];105(11):4376–80. Available from: https://pubmed.ncbi.nlm.nih.gov/18334639/
73. Miczak A, Chen B, Mun EJ, Chan W, Swenson D, Sacchettinik JC, et al. Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. 2000;261(c):735–8.
74. Mahesh PP, Retnakumar RJ, Sivakumar KC, Mundayoor S. ESAT-6 of Mycobacterium tuberculosis downregulates cofilin1 and reduces the phagosome acidification in infected macrophages. bioRxiv [Internet]. 2020 May 4 [cited 2020 Nov 24];2020.05.04.076976. Available from: https://doi.org/10.1101/2020.05.04.076976

75. Ma Y, Keil V, Sun J. Characterization of Mycobacterium tuberculosis EsxA membrane insertion: roles of N- and C-terminal flexible arms and central helix-turn-helix motif. J Biol Chem [Internet]. 2015 Mar 13 [cited 2018 Jan 15];290(11):7314–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25645924
76. Stanley SA, Cox JS. Host-pathogen interactions during Mycobacterium tuberculosis infections. Curr Top Microbiol Immunol [Internet]. 2013 [cited 2017 Oct 12];374:211–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23881288
77. Stanley SA, Raghavan S, Hwang WW, Cox JS. Acute infection and macrophage subversion by Mycobacterium tuberculosis require a specialized secretion system. Proc Natl Acad Sci U S A [Internet]. 2003 Oct 28 [cited 2020 Nov 24];100(22):13001–6. Available from: www.pnas.org
78. Rosenberg OS, Dovala D, Li X, Connolly L, Bendebury A, Finer-Moore J, et al. Substrates control multimerization and activation of the multi-domain ATPase motor of type VII secretion. Cell [Internet]. 2015 Apr 23 [cited 2020 Nov 24];161(3):501–12. Available from: /pmc/articles/PMC4409929/?report=abstract
79. Bitter W, Houben ENG, Bottai D, Brodin P, Brown EJ, Cox JS, et al. Systematic genetic nomenclature for type VII secretion systems. PLoS Pathog. 2009;5(10):8–13.
80. DiGiuseppe Champion PA, Stanley SA, Champion MM, Brown EJ, Cox JS. C-terminal signal sequence promotes virulence factor secretion in Mycobacterium tuberculosis. Science (80- ) [Internet]. 2006 Sep 15 [cited 2020 Jul 10];313(5793):1632–6. Available from: https://science.sciencemag.org/content/313/5793/1632
81. Simeone R, Bobard A, Lippmann J, Bitter W, Majlessi L, Brosch R, et al. Phagosomal rupture by Mycobacterium tuberculosis results in toxicity and host cell death. PLoS Pathog. 2012 Feb;8(2):e1002507.
82. Singh SK, Yamashita A, Gouaux E. Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. Nature [Internet]. 2007 Aug 23 [cited 2020 Nov 24];448(7156):952–6. Available from: https://pubmed.ncbi.nlm.nih.gov/17687333/

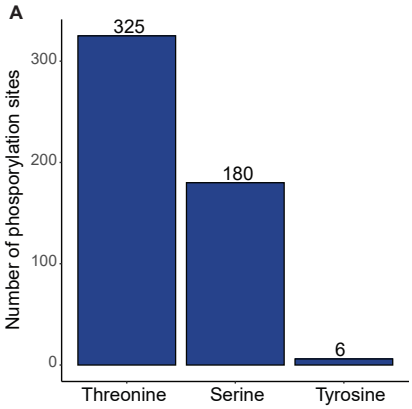
# SUPPLEMENTARY FIGURES

**Supplementary table 1:** Top 10 GO terms, associated with supplementary figure 1

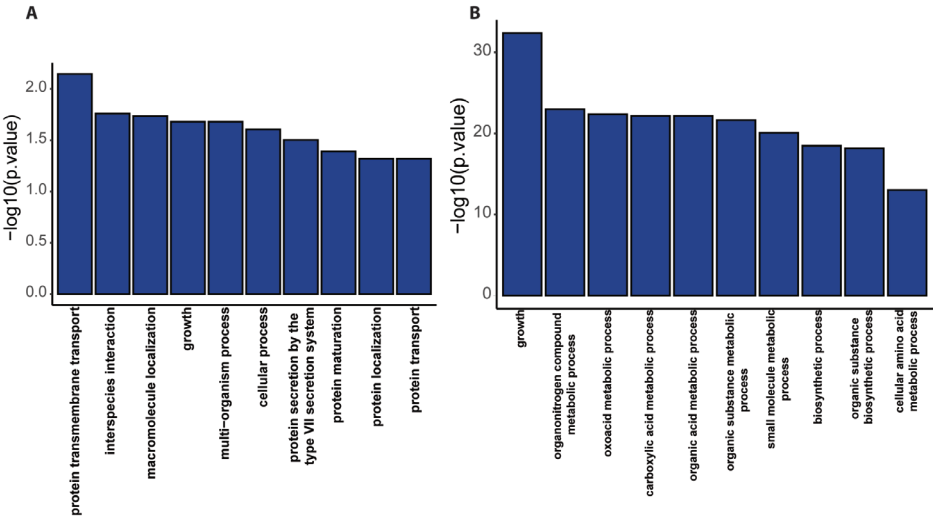
| Rank | GO ID | Term |
|---|---|---|
| 1 | GO:0019359 | nicotinamide nucleotide biosynthetic process |
| 2 | GO:0016717 | oxidoreductase activity |
| 3 | GO:0046496 | nicotinamide nucleotide metabolic process |
| 4 | GO:0019363 | pyridine nucleotide biosynthetic process |
| 5 | GO:0019362 | pyridine nucleotide metabolic process |
| 6 | GO:0009435 | NAD biosynthetic process |
| 7 | GO:0006733 | oxidoreductase activity |
| 8 | GO:0003700 | DNA-binding transcription factor activity |
| 9 | GO:0070482 | response to oxygen levels |
| 10 | GO:0072525 | pyridine-containing compound biosynthetic process |




**Supplementary figure 1: Gene ontology enrichment analysis on the down regulated proteins related to Figure 1:** Figure displays the odds ratio's as well as the top 10 most enriched terms gene ontology terms. The description of each term can be found in the table below. Terms were enriched using a hypergeometric test from all proteins depicted in Figure 1B with a log fold change below -1 and a q-value below 0.05.




**Supplementary figure 2: Metadata associated with differentially regulated phosphorylated proteins, related to Figure 2. A)** Bar plot displaying the number of phosphorylated serine's, threonine's or tyrosine's across all phosphorylated proteins. **B)** Tandem mass spectra of all regulated phosphorylated peptides available in online supplementary material. Peptides were considered regulated if the ratio of phosphorylated peptide to protein abundance was two fold or greater. Peptides were considered confident if a posterior error probability score was less than 0.01 and the localization probability was greater than 0.75.




**Supplementary figure 3: Gene ontology enrichment analysis on the protein turnover classes, related to Figure 3.** Figure displays the top 10 regulated terms as determined by hypergeometric tests from **A)** Class II or **B)** class III proteins. Classes were calculated from protein turnover data gathered at physiological pH only.

# 4

# Identification of gene fusion events in *Mycobacterium tuberculosis* that encode chimeric proteins

James Gallant[1,2]
Jomien Mouton[1]
Roy Ummels[3]
Corinne ten Hagen-Jongman[2]
Nastassja Kriel[1]
Arnab Pain[4,5]
Robin Warren[1]
Wilbert Bitter[2,3*]
Tiaan Heunis[1,6,*]
Samantha Sampson[1,*]

[1] DST/NRF Centre of Excellence in Biomedical TB research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Department of Biomedical Science, Faculty of Medicine and Health Science, Stellenbosch University, Tygerberg, 7505, Cape Town, South Africa
[2] Section Molecular Microbiology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam 1081 HZ, Amsterdam, The Netherlands
[3] Medical Microbiology and Infection Control, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam 1081 HZ, The Netherlands
[4] Biological and Environmental Sciences and Engineering (BESE) Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia.
[5] Global Station for Zoonosis Control, GI-CoRE, Hokkaido University, N20 W10 Kita-ku, Sapporo, Japan
[6] Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

Nucleic Acids Research: Genomics and Bioinformatics

DOI: 10.1093/nargab/lqaa033

## ABSTRACT

*Mycobacterium tuberculosis* is a facultative intracellular pathogen responsible for causing tuberculosis. The harsh environment in which *M. tuberculosis* survives requires this pathogen to continuously adapt in order to maintain an evolutionary advantage. However, the apparent absence of horizontal gene transfer in *M. tuberculosis* imposes restrictions in the ways by which evolution can occur. Large scale changes in the genome can be introduced through genome reduction, recombination events and structural variation. Here, we identify a functional chimeric protein in the *ppe38-71* locus, the absence of which is known to have an impact on protein secretion and virulence. To examine whether this approach was used more often by this pathogen, we further develop software that detects potential gene fusion events from multigene deletions using whole-genome sequencing data. With this software we could identify a number of other putative gene-fusion events within the genomes of *M. tuberculosis* isolates. We were able to demonstrate the expression of one of these gene fusions at the protein level using mass spectrometry. Therefore, gene fusions may provide an additional means of evolution for *M. tuberculosis* in its natural environment whereby novel chimeric proteins and functions can arise.

## INTRODUCTION

*Mycobacterium tuberculosis,* the causative agent of the disease tuberculosis, is a facultative intracellular pathogen. Adaptation to an intracellular niche is typically accompanied by reductive evolution, which favours the accumulation of pseudogenes and gene deletions. These events are prevalent in bacterial pathogens (1), mutualistic endosymbiotic bacteria (2) and are observed in obligate pathogenic mycobacteria (3–5). Reductive evolution is most commonly associated with the loss of redundant genetic material, as a pathogen or symbiont transitions towards an intracellular lifestyle (4). As an intracellular species, *M. tuberculosis* experiences a constant selective pressure in competition with the host, even though genome reduction in this pathogen has been limited. In order to maintain a competitive advantage, evolutionary changes are required. Usually, lateral evolution (horizontal gene transfer) plays an important role in bacterial adaptation to specific environments (6). Horizontal gene transfer is crucial to asexually reproductive organisms as this process provides a mechanism for the incorporation of genetic diversity in response to a dynamic environment (7, 8). This phenomenon likely occurs due to isolation of the bacilli while residing within a specialised environment such as a human macrophage. It has previously been shown that horizontal gene transfer occurs at higher frequency between closely related organisms who share the same niche (9, 10). However, horizontal gene transfer is limited or even absent in intracellular pathogens, likely due to isolation of the bacilli while residing within a specialised intracellular environment. In line with this, *M. tuberculosis* displays minimal signs of recent horizontal gene transfer events in the genome (6). *M. tuberculosis* relies on mechanisms independent of lateral evolution to acquire new material that facilitates continued evolution, such as genome recombination, gene duplication events and single nucleotide variants (11). While the contribution of single nucleotide variants in *M. tuberculosis* evolution and adaptation has been well characterised (12), the functional contribution of large scale genomic variation is largely understudied.

It was recently reported that a multi-operon deletion in the *ppe38-71* operon of *M. tuberculosis* had a major effect on the surface characteristics of this pathogen, as it resulted in the loss of all secreted Proline-Glutamic acid Polymorphic GC Rich Sequence (PE-PGRS) and PPE-Major Polymorphic Tandem Repeats (MPTR) proteins (13). This was especially striking due to the deletion occurring naturally within members of the highly successful Beijing strain family of *M. tuberculosis* (13). Of interest was the specific nature of this deletion, where the breakpoints fell within the open reading frames of distally located genes. We thus hypothesised that this type of rearrangement can result in the formation of novel chimeric proteins. However, this process will only produce a functional chimeric protein if the frame is maintained when creating gene

fusions. This phenomenon has garnered much attention in the cancer research field, where cells are prone to large scale rearrangements in the genome (14–16) and similar mechanisms have recently been demonstrated in bacteria (17, 18). We reasoned that the formation of natural chimeric proteins provides a mechanism for functional large-scale alterations in the genome in the absence of horizontal gene transfer.

We have previously used comparative genomics in conjunction with discovery-based proteomics to create custom proteome search databases for analysing strain-specific features in clinical *M. tuberculosis* isolates (19). Here, we expand on our approach and extend it to naturally occurring gene fusions by designing and implementing custom software to identify fusions in the genomes of *M. tuberculosis* clinical isolates. In addition, we were able to confirm the expression of fusion proteins using tandem mass spectrometry. We demonstrate an additional means for *M. tuberculosis* evolution, which is likely applicable to the adaptation of other intracellular pathogens as well.

## MATERIALS AND METHODS

### Bacterial culture

*Mycobacterium tuberculosis* CDC1551 and H37Rv were used as wild type strains; *Δppe38-71* and *Δppe38-71::*pMVHSP60-*ppe38-71* (complemented strain) were generated previously from *M. tuberculosis* CDC1551 as the parental strain (13). Clinical isolates of *M. tuberculosis* used in this study were obtained from the South African Western Cape region. Procurement of clinical isolates were approved by the Stellenbosch University Health Research Ethics Committee (approval number: N10/04/126). Samples were de-identified and not linked to any patient information. Samples obtained from patients were initially cultured in mycobacterial growth indicator tubes (MGIT) at 37°C until growth was detected by means of a BACTEC 960 broth culture system (BD Bioscience, NJ, USA). Positive MGIT cultures were centrifuged, stored as glycerol cryobead stocks in the Division of Molecular Biology and Human Genetics strain bank as bacterial seed lots or represented by genomic DNA.

For further experimentation, mycobacterial cultures were either cultured in modified Sauton's media (0.4% L-asparagine, 0.4% glucose, 0.2% citric acid, 0.05% monopotassium phosphate, 0.05% magnesium sulphate, 0.005% ferric ammonium citrate, 0.1 ml of 1% zinc sulphate and 0.05% Tween-80, pH 7.0) or Middlebrook 7H9 media supplemented with an oleic acid, albumin, dextrose and catalase (OADC) mix (BD bioscience, NJ, USA), 0.5% glycerol and 0.05% Tween-80. Mycobacterial cultures were grown

without shaking at 37°C until harvesting. Long-term storage of mycobacterial cultures were done at -80°C in 20% glycerol.

*Escherichia coli* Top10F' cells were cultured in lysogeny broth (LB) (1% Bacto-tryptone, 0.5% Bacto-yeast extract, 1% Sodium chloride) media with shaking at 37°C, or kept at -80°C in 20% glycerol until use.

## Mycobacterial strain selection

Twenty-one *M. tuberculosis* strains were selected for initial analysis (Table S1); 17 of these were genotyped based on spoligotyping and restriction fragment length polymorphism (RFLP) was used to stratify whether the lineage 2 strains were typical or atypical as previously described (20, 21), while the remaining 4 strains (HN878, S1945, S2135, S2701) were genotyped in previous studies (13, 22). These specific strains were chosen based on their genetically diverse profile from two commonly occurring mycobacterial lineages in the South African Western Cape region and formed the basis of our study (Additional file 1: Fig. S1A). An additional 159 clinical isolates were chosen from an in-house mycobacterial genome sequence repository for a total of 180 *M. tuberculosis* clinical isolates (Additional file 1: Fig. S1B). In total, 90 independently isolated strains from each lineage was procured to gain greater statistical power during computational analysis. These samples have a genetic bottleneck due to the limited sites where *M. tuberculosis* isolates can be sampled. To account for this, the lineages were determined *in silico* using TBprofiler to avoid including multiple strains that represent a single sub-lineages (23).

## DNA extraction and whole genome sequencing

Genomic DNA was extracted as previously described (24). The DNA library was sequenced on an Illumina HiSeq2000 platform (Illumina, inc, CA, USA). Briefly, one microgram of DNA was used for library preparation using the Nextera DNA sample preparation kit according to the manufacturer's instructions. Paired-end reads were sequenced using approximately 500 bp fragment sizes. FastQ files for previously sequenced clinical isolates used in this study were obtained from the Stellenbosch University mycobacterial genome sequence repository.

## Construction of genomics pipeline

We have developed an in-house automated data analysis pipeline using the bash scripting language on the Ubuntu distribution of Linux for the detection of large deletions and potential gene fusions. Illumina FastQ files were processed in accordance with the genome analysis toolkit (GATK) best practices guide using either *M. tuberculosis* H37Rv (NC_000962.3) or *M. tuberculosis* CDC1551 (AE000516.2 ) as the reference strain (25, 26).

FastQ files were trimmed and aligned using BWA and NovoAlign software (27). The resulting binary alignment map (BAM) files were used for single nucleotide variation (SNV) detection using GATK and SAMtools (25, 28). Automatic *in silico* lineage typing was determined by piping raw FastQ files to TBprofiler (23). Detection of deletions by read-pair and split reads was done using Delly and Lumpy while Bedtools was used to detect deletions based on zero coverage (29–31).

## Construction of gene fusion identification pipeline

To identify gene fusions, we first extracted a list of all the structural variants (SVs) found within the genome. All SVs used for high throughput gene fusion detection were obtained from either Delly (29) or Lumpy (30). These include insertions, deletions, duplications, inversions and translocations. This list of SVs was further filtered for deletions, as these had the highest likelihood of recombining into gene fusions. The breakpoints of these deletions are subsequently annotated against the reference database with either the affected gene or the intergenic region. Using this information, we applied a set of filtering parameters to identify a list of putative gene fusions.

Firstly, we excluded any deletions that did not span multiple genes. If this criterion was passed, the deletion breakpoints had to fall within open reading frames of genes on each side of the breakpoints, and these genes had to be in the same orientation. If all these criteria were met, 2000 bp flanking each side of the breakpoint was extracted from the BAM files and converted to FastQ format. This format was used to determine optimal kmers using kmergenie and *de novo* assembled using SOAPdenovo2 which allows for precise breakpoint detection (32, 33). The resulting contigs were ordered against the corresponding truncated reference sequence of *M. tuberculosis* H37Rv using ABACAS which results in a consensus sequence (34). Post-pipeline analysis included manually inspecting the potential fusion sequence for open reading frames by searching in six frames using NCBI ORF finder and translated to protein sequence (35). These amino acid sequence from each potential fusion protein was cross referenced to both parent amino acid sequences. If a match occurs, and no insertion sequences are present in the genomic region, the fusion amino acid sequences were added to a protein FASTA database. This database was ultimately used for downstream peptide identification by mass spectrometry.

## Harvesting of whole cell lysates and supernatants

*M. tuberculosis* isolates were cultured in modified Sauton's media supplemented with Tween-80 and allowed to propagate for seven days at 37°C until an $OD_{600}$ of 1.0 without shaking. *M. tuberculosis Δppe38-71* and the complemented strain strains were supplemented with either hygromycin (50 μg/ml) or kanamycin (25 μg/ml) and hygromycin

(50 µg/ml), respectively. Antibiotics were omitted during growth for all clinical isolates. Cultures were washed three times with phosphate-buffered saline (PBS) to remove residual Tween-80 and inoculated in modified Sauton's media without Tween-80 at an $OD_{600}$ of 0.05. Bacterial cultures were allowed to propagate for an additional seven days. Not all clinical isolates defined in supplementary table 1 were able to grow in the modified Sauton's media, likely due to the lack of OADC, and were therefore omitted when determining PE-PGRS secretion. The bacterial cells and supernatant were separated by centrifugation (4000 rpm, 10 minutes) followed by resuspension of the cell pellet in lysis buffer (8M urea in 100 mM tetraethylammonium bromide (TEAB), 5 mM tris(2-carboxyethyl)phosphine (TCEP), Benzonase, Roche cOmplete™ EDTA free cocktail tablets). Cell-free supernatants were filter sterilised using a 0.22 µM steriflip filter unit (Merck Milipore, MA, USA). Whole-cell lysates were prepared by bead beating the resuspended cell pellet (20s cycles with 20 seconds on ice for a total of 8 cycles), followed by clarification (14 000 rpm, 10 minutes, 4°C). Sterilised cell-free supernatants were concentrated with Amicon ultra-15 kDa spin filters (Merck Milipore, MA, USA), to approximately 200µl by centrifugation (14 000 rpm, 4°C). Concentrated cell-free supernatants were precipitated overnight at -20°C using four volumes of ice-cold acetone. Protein content was quantified using a modified version of the Bradford protein assay (36).



## Sample preparation and liquid chromatography tandem mass spectrometry

Whole-cell lysates of *M. tuberculosis* clinical isolates S507, S5527 and S5218 were prepared in biological triplicate as written above and processed for liquid chromatography tandem mass spectrometry (LC-MS/MS). Whole cell lysates were digested to peptides for shotgun proteomics following an in-solution digestion protocol. Briefly, equal concentration of proteins resuspended in urea buffer (8M urea in 100 mM TEAB) (Sigma Aldrich, MO, USA) were reduced with 5 mM TCEP (Sigma-Aldrich, MO, USA) and alkylated with 5.5 mM iodoacetamide (Sigma-Aldrich, MO, USA) in the dark for 1 hour respectively. The protein solution was diluted with four volumes 50 mM TEAB (Sigma-Aldrich, MO, USA) to contain a concentration of less than 2M urea. Proteins were digested to peptides by addition of a 1:50 ratio of sequencing grade modified trypsin (Promega, WI, USA) to total protein content and incubated in a humidified chamber at 37°C for 18 hours. The resulting peptide mix was dried using a desiccator and resuspended in 2% acetonitrile supplemented with 0.1% formic acid. Peptides were desalted using stop and go extraction (STAGE) tips as described previously (37). The desalted peptides were dried in a desiccator and resuspended in 2% acetonitrile supplemented with 0.1% formic acid before mass spectrometry analysis.

A total of 1 µg of peptides from each sample was analysed, independently, on an Orbitrap Fusion Tribid mass spectrometer (Thermo Fisher Scientific, MA, USA) connected to a Thermo Scientific UltiMate 3000 RSLCnano system (Thermo Fisher Scientific, MA, USA). Peptides were injected on a PepMap C18 LC pre-column (300 µm ID x 5 mm, 5µm, 300 Å) followed by separation on an analytical column packed with C18 Aeris peptide 3.6 µM beads at a 500 nL/min flow rate. Solvent A was water containing 0.1% formic acid, and solvent B was 100% acetonitrile containing 0.1% formic acid. Peptides were separated as follows: solvent A was maintained at 2% followed by an increase to 7.5% in 5 minutes, 7.5% to 25% in 45 minutes, 25% to 45% in 15 minutes, 45% to 80% in 0.1 minutes, maintained at 80% for 10 minutes, followed by a decrease to 2% in 0.1 min and equilibration at 2% for 10 min. The Orbitrap Fusion was operated in positive-ion data-dependent mode and precursor ions (MS1) were detected in the Orbitrap with a nominal resolution of 120 000 at 200 m/z. An automatic gain control (AGC) target of $5 \times 10^5$ and an ion injection time of 50 ms was used. The most intense ions above a threshold of 5 x $10^3$ were selected for high-energy collision dissociation (HCD) at a normalised collision energy of 32.5%. The fragmented ions were analysed in the Orbitrap (MS2) at a resolution of 15 000 at 200 m/z. The AGC target was set to $1 \times 10^4$ and a 45 ms injection time was allowed during MS2 analysis. The number of MS2 events between MS1 scans was determined on-the-fly to maintain a 3 s fixed duty cycle. Dynamic exclusion of ions within a ± 10 ppm m/z window was implemented using a 30 s exclusion duration. An electrospray voltage of 2.0 kV and capillary temperature of 275°C, with no sheath and auxiliary gas flow, was used.

## Processing of LC-MS/MS data

A custom-made *M. tuberculosis* reference proteome was generated as described in the whole genome sequencing analysis section, specifically by addition of the putative chimeras to the reference proteome. Amino acid sequences representing potential gene fusions were added to the reference proteome (UniProt accession: UP000001584) from *M. tuberculosis* H37Rv (downloaded from UniProt 10 April 2018). This custom database was used for all mass spectrometry searches and contained for a total of 4002 entries which includes the potential gene fusions. MaxQuant (version 1.6.3.4) was used to analyse raw files obtained from LC-MS/MS. The Andromeda search algorithm, integrated in MaxQuant, was used for peptide and protein identification, using default parameters (38, 39). Carbidomethylation of cysteine was chosen as fixed modification and oxidation of methionine as well as N-terminal acetylation was chosen as variable modifications. Enzyme specificity was set as Trypsin/P and a maximum of two missed cleavages was allowed.

## PPE38 C-terminal domain cloning, expression and purification

Genomic DNA was extracted from *M. tuberculosis* CDC1551 as described above and the region encoding the PPE38 C-terminal target was amplified using primers, PPE38_AB_F; CCG CGA CGT GCT AGC ATG GCG GTG GAG GGG GTG CCG GC and PPE38_AB_R; TCA CAG GTC AAG CTT CTA CGC CGA CAT CCC CGC ACC CA with Phusion hotstart 2x master mix polymerase (NEB, MA, USA). The resulting amplicon was cloned into pIBA (40), using an In-Fusion cloning reaction as described by the manufacturer (Takara Bio, Japan), in *E. coli* Top10F' cells. Transformed cells were cultured on LB-agar supplemented with ampicillin (100 µg/ml). The presence of the *ppe38* insert was verified by restriction digest and Sanger sequencing. For recombinant protein expression, transformed *E. coli* Top10F' was cultured in LB broth supplemented with 100 µg/ml ampicillin. Expression was induced with 0.2 µg/ml anhydrotetracycline (Sigma Aldrich, MO, USA) at 37°C with shaking at 200 rpm when *E. coli* cultures reached an $OD_{600}$ of 0.3 for 2 hours or until an $OD_{600}$ of 1.5. Cultures at $OD_{600}$ of 1.5 were subsequently harvested by centrifugation (14 000 rpm, 10 minutes, 4°C) and suspended in ice-cold lysis buffer (10 mM Tris–HCl pH 8.0, 1 mM EDTA, 5 µg/ml lysozyme), followed by incubation at 37°C for 1 hour. Cells were further disrupted by sonication (Branson sonifier 250) and the lysate was centrifuged (12 000 rpm, 15 minutes, 4°C) to sediment inclusion bodies. Contaminating factors were removed by resuspending the pellet in sonication buffer (10 mM Tris–HCl pH8, 1 mM EDTA) and disrupted by sonication, followed by addition of an equal volume triton wash buffer (10 mM Tris–HCl pH 8.0, 1 mM EDTA, 2% Triton X-100) and incubation for 1 hour at room temperature. Inclusion bodies were collected by centrifugation (12 000 rpm, 15 minutes) and suspended in sonication buffer followed by another round of sonication. An equal volume of urea wash buffer (10 mM Tris–HCl pH 8.0, 2 M urea) was added followed by an hour incubation at 37°C. After incubation, the inclusion bodies were harvested by centrifugation (123 000 rpm, 15 minutes) followed by resuspension (sonication buffer) and sonication. An equal volume of high salt wash buffer (10 mM Tris–HCl pH 8.0, 2 M NaCl) was added to the resulting suspension and inclusion bodies were harvested by centrifugation (12 000 rpm, 25 minutes). Another round of sonication and centrifugation followed this step and the final inclusion body-containing suspension was stored in PBS containing 15% glycerol.

Inclusion bodies were separated by 10% SDS-PAGE (1x TGS buffer, 100 volts, 1 hour) and visualised with coomassie brilliant blue. These inclusion bodies were used for immunisation in rabbits performed by Innovagen (Sweden) as detailed in their custom polyclonal rabbit IgG service. Briefly, initial immunisation was done with the inclusion body suspension containing the PPE38 antigen with Freund's complete adjuvant. This was followed by booster immunisations at weeks 3, 5, 8 and 11 with Freund's incomplete adjuvant. Immune serum used for this study was collected at 13 weeks after im-

munisation. The resulting anti-serum from the immunised rabbit was used for western blotting.

### Western blotting

Concentrated supernatants from *M. tuberculosis* CDC1551, *Δppe38-71*, the *ppe38-71* complemented strain, S507, S3651, S1453, S3760, S4437, S3839, S5218, H37Rv, S4570, S1116, S2701 and S2135 were separated by SDS-PAGE (12% resolving gel, 100V, 1 hour) and transferred to a nitrocellulose membrane. The membrane was blocked with 1% milk powder (Sigma-Aldrich, MO, USA) for 1 hour and probed with either mouse monoclonal α-PGRS (1:5000) (41), and α-EsxA (1:500) (42) or rabbit polyclonal α-PPE38 (1:1000) (this study) overnight at 4°C. Primary antibodies were removed by washing with Tris-buffered saline supplemented with Tween (20 mM Tris-HCL pH 7.5; 150 mM NaCl; 0.1% Tween-20) followed by probing with either goat anti-mouse or goat anti-rabbit horse radish peroxidase conjugated secondary antibody for one hour. Probed nitrocellulose membranes were visualised using a ChemiDoc Gel Imaging System (Bio-rad, CA, USA). When required, membranes were stripped using mild stripping buffer (1.5% glycine, 0.1% SDS, 1% Tween-20, pH 2.2).

### PCR and Sanger sequencing

Polymerase chain reaction of the Rv2623-Rv2628 region was performed using Phusion polymerase (New England Biolabs, MA,USA) as indicated by the manufacturer with the addition of 2% DMSO. Products from *M. tuberculosis* H37Rv and clinical isolate S507 were sequenced by capillary electrophoresis to identify and confirm the fusion junction using the following primers: Rv2623_F; CCA TTG TCG CGC ACA AAC and Rv2628_R; GTG GCA TGG CCA TGT CTT CTA.

### Statistical analysis

All statistical tests performed downstream using Delly and Lumpy-SV outputs were implemented in the R programming language version 3.6.2. The data presented in Figure 1A was analysed as a contingency table using Fisher's exact test with a p-value cut off set at 0.05.

## RESULTS

### *ppe38-71* deletions are more prevalent in lineage 2 isolates of *M. tuberculosis*.

A previous study in our group has shown that deletions within the *ppe38-71* locus in certain strains of *M. tuberculosis* can result in a block of PE-PGRS secretion and is as-

sociated with increased in virulence (13). An earlier study already indicated that the *ppe38-71* operon is prone to recombination with multiple variations (43). However, because of the highly repetitive nature and high GC content of this region, these deletions are difficult to identify. In this study, we further investigated the genetic relationship between *ppe38-71* deletions in various lineages of *M. tuberculosis* and the effect of this mutation on PE-PGRS protein secretion. Twenty-one well characterised and genetically diverse clinical isolates representing lineage 2 and lineage 4 were used as the initial screening group (Additional file 1: Fig. S1A, Supplementary table 1).



A custom Illumina data analysis pipeline was constructed, which used two split read callers, a coverage-based approach and a targeted approach, to find both known and unknown deletions (Additional file 1: Fig S2). To address the variability within the region, two genes that fall between *ppe38* and *ppe71* (*mt2420* and *mt2421*) were examined. These genes are unaffected by the insertions of IS6110 and reads mapping to this region were therefore used as an indicator for a full *ppe38-71* operon (Additional file 1: Fig. S3). Using these criteria, *ppe38-71* deletions were detected in both lineage 2 and lineage 4. The majority of the lineage 2 strains (Fig. 1A) had *ppe38-71* deletions while the converse was observed for lineage 4 strains (Fig. 1B). This suggested a greater prevalence of this deletion, and by association a predicted lack of PE-PGRS secretion, in lineage 2 isolates. However, the members of both lineages are sampled within the same local geographical region and thus provide a bias within each distinct lineage. To detect whether this deletion was indeed more prominent in *M. tuberculosis* lineage 2 as isolates compared to lineage 4, 90 members from each of the two lineages were further screened for a total of 180 genomes (Additional file 1: Fig. S1B, Fig. S4 A1-A6). Using the same metric to detect this deletion as described above, we found a 74% occurrence of the *ppe38-71* deletion in lineage 2 and a 17% occurrence in lineage 4. Thus indicating a significantly higher prevalence (p-value < 2.2e-16) of *ppe38-71* deletions in the lineage 2 isolates. (Fig. 1C).

Previous publications demonstrated that breakpoints occur within *ppe38* and *ppe71* coding regions, effectively truncating the coding sequences of these genes and abolishing PE-PGRS secretion (13, 43). This phenotype was tested in selected lineage 2 and lineage 4 clinical isolates by immunoblotting with a monoclonal antibody that specifically recognises repeat domains (41). No PE-PGRS secretion was observed in members of lineage 2 that also had a *ppe38-71* deletion, as determined by whole genome sequencing (Fig. 1D). Interestingly, PE-PGRS secretion was observed in S3651, which also has a *ppe38-71* deletion as determined by whole genome sequencing (Fig. 1D).

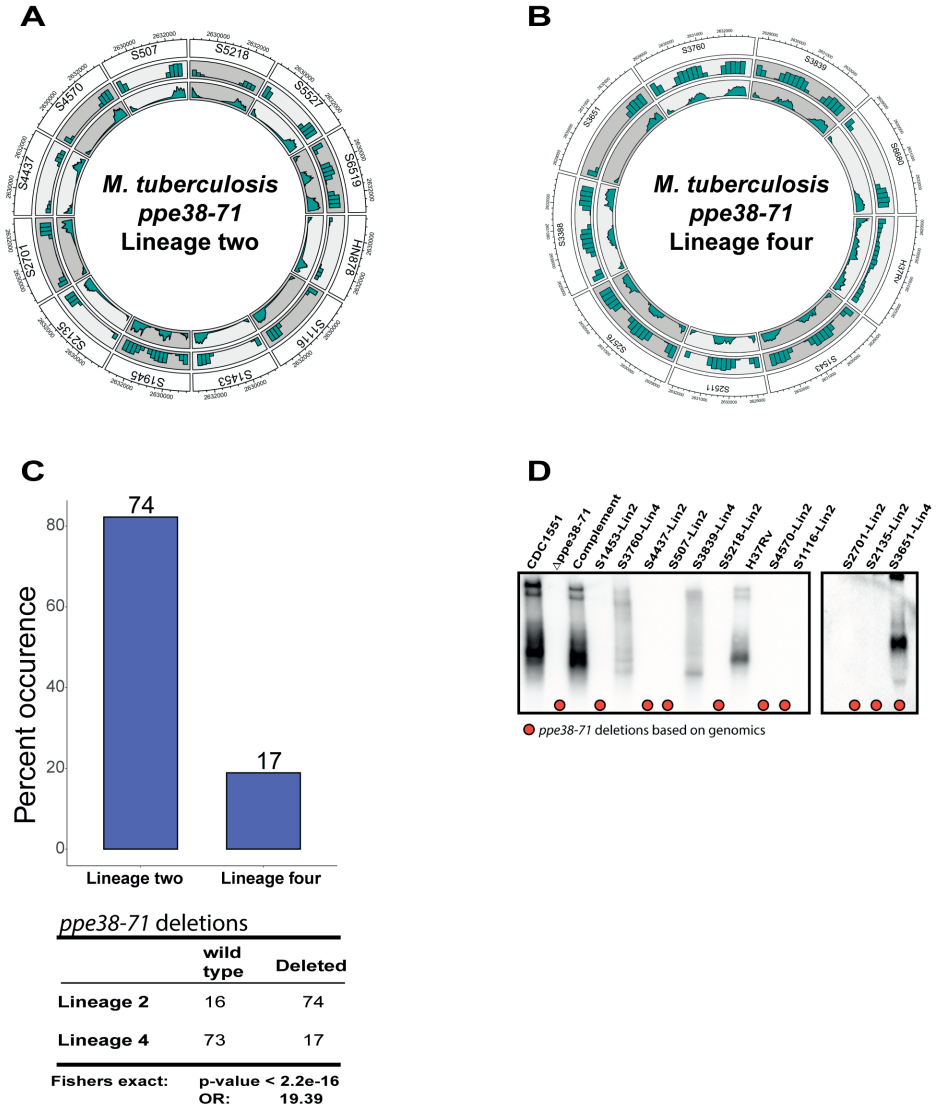



**Figure 1: Deletions in the *ppe38-71* operon are more prevalent in *M. tuberculosis* lineage 2 isolates.** Deletions are represented as Circos plots displaying whole genome sequence alignment by genomic coordinates (outer track), strain (first track), read density (middle track) and average coverage (inner track) for clinical isolates of *M. tuberculosis* representing **A)** lineage 2 and **B)** lineage 4. Averages were determined per strain and only within the genomic coordinates spanning the deletion. **C)** Occurrence of *ppe38-71* deletion as a percent calculated from each sub-population of lineage 2 and lineage 4 representing 90 clinical isolates, respectively. Deletions were determined by inspecting coverage in the *ppe38-71* operon. **D)** Western blot of cell-free supernatants obtained from *M. tuberculosis* clinical isolates, *M. tuberculosis* CDC1551 as well as control strains targeting PE-PGRS proteins. Red dots indicate clinical isolates that were predicted by whole genome sequencing to have a *ppe38-71* deletion.

In the ancestral operon, the *ppe38* and *ppe71* genes are nearly identical copies of one another separated by two *esx* genes (43). It has been suggested that deletion of the intervening sequence can result in a fusion protein (43). We thus reasoned that the presence of PE-PGRS proteins in S3651 culture supernatants could be due to the internal deletion causing the open reading frames of ppe38 and *ppe71* to form a chimeric protein with a similar function to PPE38.

## *PPE38-71* deletion creates a natural chimera in clinical isolate S3651

We investigated different configurations of the *ppe38-71* operon in more detail to determine whether a functional *ppe38/71* chimeric protein was indeed present in the lineage 4 isolate, S3651. For this, we compared S3651 to S507 as an example of a lineage 2 isolate with a *ppe38-71* deletion, S3388 as an example of a lineage 4 isolate with a full-length *ppe38-71* operon, and *M. tuberculosis* H37Rv which also has a full-length *ppe38-71* operon (Fig. 2A).



As transposon insertions have previously been found in the *ppe38-71* operon (13), we reasoned that the presence of a transposon will disrupt the open reading frames in cases where PE-PGRS secretion is abolished. *De novo* assemblies were used to target the *ppe38-71* operon in S507 and S3651, which represent a PE-PGRS secretion negative and positive phenotype, respectively. This was accomplished by extracting aligned reads for approximately 2000bp upstream and downstream of the *ppe38* and *ppe71* break points. Targeted *de novo* assembly revealed that S507 (lineage 2) has an insertion sequence present between *ppe38* and *ppe71,* which disrupts the reading frames (Fig. 2B, Additional file 1: Fig. S5A). However, *de novo* assemblies of S3651 indicated breakpoints causing the open reading frames of *ppe38* and *ppe71* to fuse and create a *ppe38/71* gene fusion (Fig. 2C). The *ppe38/71* fusion is created through an out-of-frame deletion spanning the *ppe38* and *ppe71* operon, situated on the reverse strand, with breakpoints occurring within the open reading frame of each respective protein. Specifically, a cytosine-cytosine pair remaining on the *ppe71* breakpoint combines with a thymine at the *ppe38* breakpoint generating a cytosine-cytosine-thymine codon (Additional file 1: Fig. S5B). This recombination reinstates the proline originally encoded by cytosine-cytosine-guanine in *ppe71*, thereby creating the amino acid sequence S G P I A S reading from the N-terminal of PPE71.

Taking into account the secretion of PE-PGRS proteins found in S3651 (Fig. 1D) as well as the predicted reformation of a distinct, yet functional equivalent of *ppe38* in S3651 (Fig. 2C) we reasoned that this gene fusion is likely responsible for the PE-PGRS secretion phenotype. To detect expression of *ppe38,* we generated a rabbit polyclonal antibody against the C-terminal domain of *M. tuberculosis* PPE38, which was used to immunise

rabbits. Rabbit anti-serum was first tested for reactivity against PPE38 (Additional file 1: Fig. S6A, Fig. S6B). Western blots against *M. tuberculosis* supernatants revealed the presence of PPE38 epitopes in *M. tuberculosis* CDC1551, the *ppe38-71* complemented strain as well as *M. tuberculosis* S3651. As expected, PPE38 was not detected in S507 and *Δppe38-71* strains (Fig. 2D). Interestingly, two bands were detected when probing for PPE38, and both bands were absent in both S507 and *M. tuberculosis Δppe38-71* (Fig. 1D,

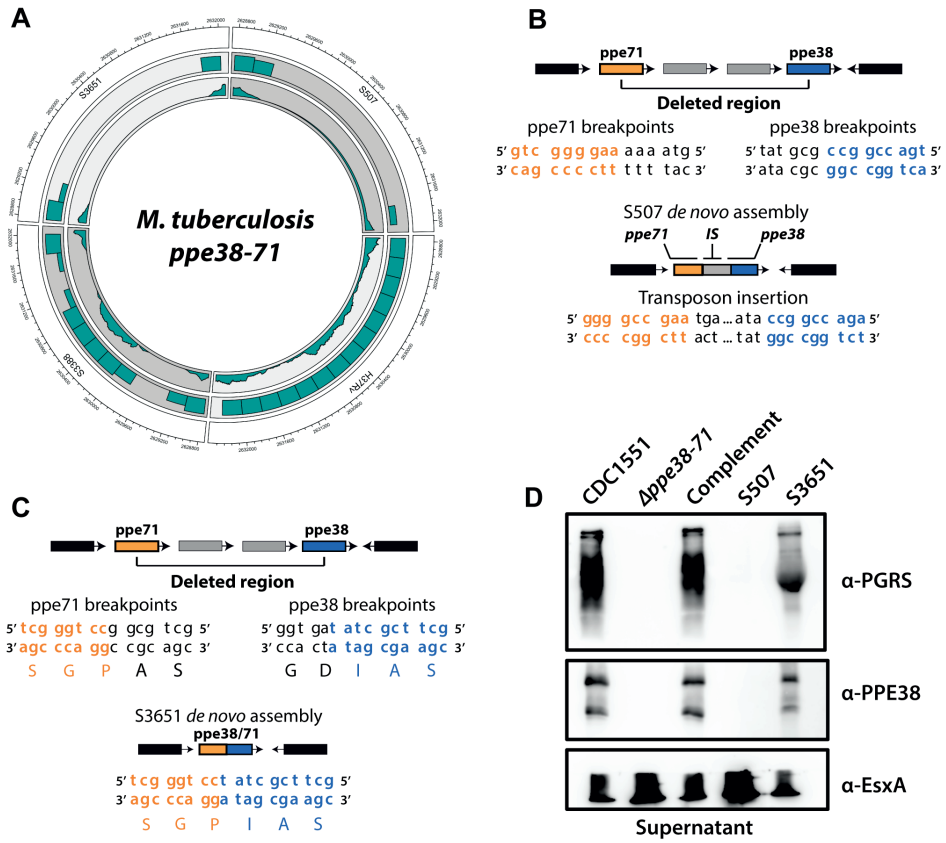



**Figure 2: *M. tuberculosis* lineage 4 strain with a *ppe38-71* deletion produces a functional chimeric protein. A)** Specific clinical isolates used for further investigation of the *ppe38-71* operon. *M. tuberculosis* H37Rv was used as a control for the NGS and aligned to CDC1551, full length *ppe38-71* is detected in contrast to the published reference. S3651 and S507 represents clinical isolates with a *ppe38* deletion and S3388 a clinical isolate without. **B)** Schematic representation of targeted *de novo* assembly and contig ordering of S507 indicating a transposon insertion between *ppe71* and *ppe38*, thereby disrupting the *ppe71* reading frame. **C)** Schematic representation of targeted *de novo* assembly in the *ppe38-71* operon of S3651. No transposon insertion was found and the reading frame of *ppe71* is intact causing a gene fusion. **D)** Western blot of *CDC1551* reference, *Δppe38-71, ppe38*-71 complemented strain, S507 and S3651 cell-free supernatant probed for PE-PGRS proteins, PPE38 and ESAT-6 as the loading control. S3651 secreted PE-PGRS proteins and expressed PPE38 similarly to the wild type

Additional file 1: Fig. S6B). This may occur due to cleavage PPE38 through a protease, possibly PecA, which has been shown to cleave PE-PGRS proteins (44). Staining with anti-serum directed against PE-PGRS proteins resulted in a similar pattern, providing evidence that the production of PPE38 and secretion of PE-PGRS are linked and that the chimera formed in S3651 is associated with the PE-PGRS secretion.

## Detecting multigene deletions in *M. tuberculosis* clinical isolates

The presence of functional gene fusions in *M. tuberculosis* is an intriguing finding, which in the absence of horizontal gene transfer, may provide an alternate method for evolutionary adaptation. To find additional chimeric proteins which has the potential to form gene fusions, we searched for genomic features likely to result in a gene fusion event. This was automated and coupled with *de novo* assembly, which was used to generate consensus sequences surrounding multigene deletions (Additional file 1: Fig. S7). This approach served as a first-pass method to identify multigene deletions present within coding sequences and is thus a starting point to identify potential gene fusions. This algorithm was incorporated into our existing software and used to screen clinical *M. tuberculosis* isolates for possible gene fusions.



We screened the genomes of the same 180 clinical isolates mentioned above for multigene deletions that can form potential gene fusions and compared this to all the SVs detected. The occurrence of gene fusions represented a low proportion of the total structural variation (Fig. 3A). Furthermore, the SV count had more variability across the two lineages with some strains containing more than 100 SVs. This was observed in both lineages, however, this increase in the total number of SVs did not affect the distribution of potential gene fusions (Fig. 3A). Thus, formation of gene fusions seem to be a rare phenomenon likely due to the inherent randomness in genetic variation and the precise breakpoints required to maintain a reading frame. Once this process concludes and a chimeric protein is formed, the new protein has to be favoured by natural selection to allow for propagation of the new feature within the population. It therefore stands to reason that formation of fused genetic elements that yield proteins are less likely to occur compared to pseudogene formation, full gene formations or out-of-frame deletions. The highest occurrence of multigene deletions was found in genes encoding PE/PPE proteins, however due problematic read alignment in these regions (very high GC content of more than 75% and multiple repeats) it is difficult to determine with confidence that these are indeed true gene fusion events (Fig. 3B). We therefore discarded multigene deletions occurring within genes annotated as PE/PPE proteins during further analysis. This resulted in the identification of 21 multigene deletions, two of which were found within both lineages. Multigene deletions that resulted in fused open reading frames from the genomic data were further manually inspected by

six-frame translation of the fusion gene. If the six-frame translation indicated a single protein with domains from both parent genes, we annotated this as fused (Fig. 3C). We could detect the $RD^{RIO}$ deletion which forms a fused reading frame in Rv3346c/55c (Fig. 3C) (45) . However, the most prevalent multigene deletion that has fused open reading frames in our sample set was represented by the RD105 deletion (*Rv0071/74*) that occurs in lineage 2 isolates (46) (Fig. 3C).

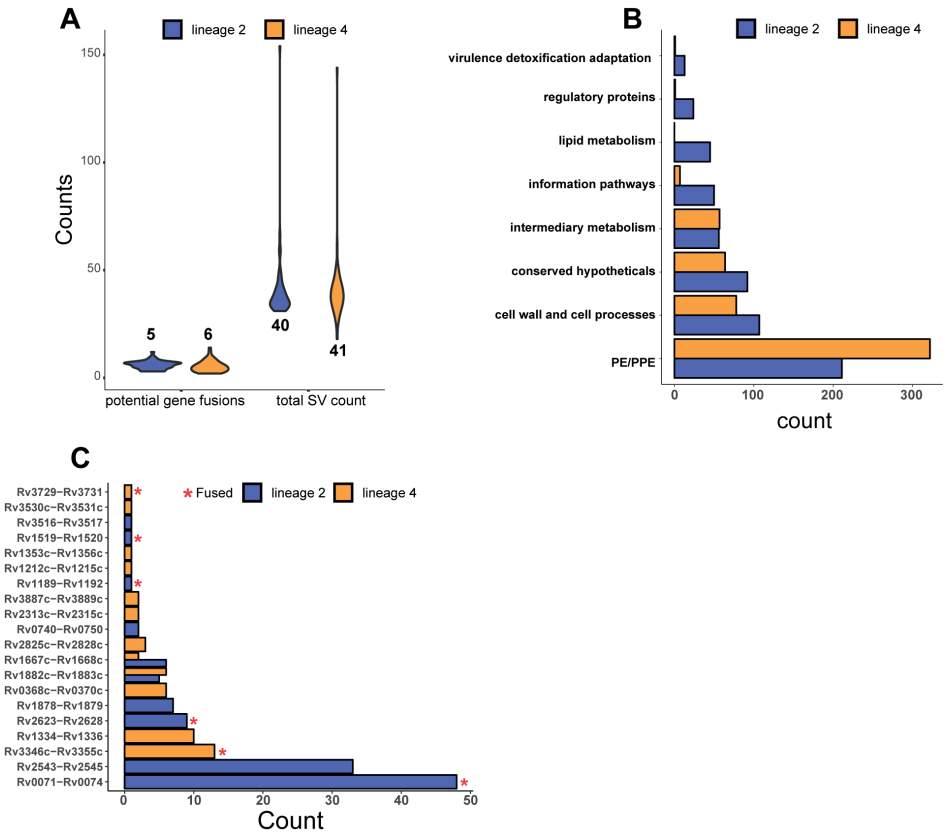



**Figure 3: Systematic detection of potential fusion events from *M. tuberculosis* large sequence polymorphisms. A)** The distribution of multigene deletions that fall within an open reading frame and have the same orientation as predicted by our software compared to the distribution of structural variants (SV) found across 180 isolates of lineage 2 and lineage 4. The numbers on the graph represent the mean of the distribution. **B)** Most abundant annotations associated with potential gene fusions across all clinical isolates and separated by lineage. Annotation terms were sourced from the Mycobrowser functional categories. PE/PPE proteins constituted the majority of identifications associated with gene fusions. Potential gene fusions falling in this category were removed from further consideration as alignment failures in these areas are highly prevalent. **C)** Occurrence of specific multigene deletions that fall within open reading frames. Each of these were manually inspected for closed reading frames and annotated as either fused or truncated.

## Genetic evidence for gene fusions in mycobacteria

Three candidate gene fusions were observed at relatively high frequency within the 180 *M. tuberculosis* genomes, compared to the other gene fusions (Fig. 3C). We opted to focus on *Rv0071/*74 and *Rv2623/*28 for further investigation, while *Rv3346c/55c* was disregarded as this specific deletion was not found within the 21 well characterised clinical isolates and has been described before (45). The *Rv2623-Rv2628* genomic region is part of the *M. tuberculosis* dormancy regulon and has not yet been reported as deleted within clinical isolates of *M. tuberculosis.* This region is deleted in lineage 2 members, S507 and S5527, where the deletion completely removes the genes *Rv2624-Rv2627* and truncates *Rv2623* at the 5' end, coding for the C-terminal domain, and *Rv2628* at the 3' end (Fig. 4A). The RD105 deletion was also detected as a gene fusion using our pipeline and recently shown to indeed form a chimeric protein (Fig. 4B) (18). This deletion is prevalent in the lineage 2 isolates and is used as a marker for sub-lineage speciation within *M. tuberculosis* (46). The protein encoded by *Rv0071* is annotated as a potential maturase and the *Rv0074* gene product is of unknown function and is localized in the membrane of *M. tuberculosis,* as determined by mass spectrometry analysis (47).



The Rv2623/28 putative gene fusion, identified by our method, has not yet been reported as a gene fusion or indeed a deletion in the mycobacteria. We therefore characterised this feature further by verifying the presence of this deletion as well as the specific break points using polymerase chain reaction and capillary electrophoresis sequencing. First, the deletion was confirmed in S507 and S5527 identifying a band at the 700 bp range using primers flanking *Rv2623* and *Rv2628*, compared to a band of ~7000 bp in S5218 and H37Rv which have an intact operon (Fig. 4C). Next the 700 bp and ~7000 bp band corresponding to S507 and S5218 was sequenced, respectively. The base pair sequence corresponding to wild type *Rv2623* and the *Rv2623/28* fusion was discernible and corresponded to the sequence prediction from our *de novo* assemblies of this genomic region (Fig. 4D). Based on these observations, the Rv2623/28 remains intact and should transcribe a hybrid protein under the Rv2623 promoter.

## Gene fusions form chimeric proteins.

*M. tuberculosis* follows a reductive evolutionary path and consequently has a number of pseudogenes (48). Although the candidates found in our genetic screens may have the genotypic characteristics of a chimeric protein, these may not lead directly to proteins. The genetic features may thus be present within the genome, not as functional chimeras but rather pseudogenes.

We used mass spectrometry analysis to investigate whether the putative gene fusions found in our genetic screens are expressed by *M. tuberculosis* to form stable chimeric

proteins. To identify chimeric proteins, the translated sequences obtained from the *de novo* assemblies of *ppe38/71;* Rv2623/28 and Rv0071/74 were added to the *M. tuberculosis* H37Rv protein database (UP000001584). The *pks15/1* gene is found in W-Beijing strains of *M. tuberculosis* and is comprised of fused *pks15* and *pks1* genes (22, 49). This gene was not detected in our genetic screens of W-Beijing strains, likely due to the close proximity of the breakpoints causing discordant read pairs to be missed. We therefore added this chimeric protein sequence as well, as it has previously been shown to combine in a manner similar to what we predict for gene fusions (22, 49). This database was used to search tandem mass spectra from S507, S5527, S5218 generated in this study as well as S3651 which is publicly available and generated from a previous study (19).

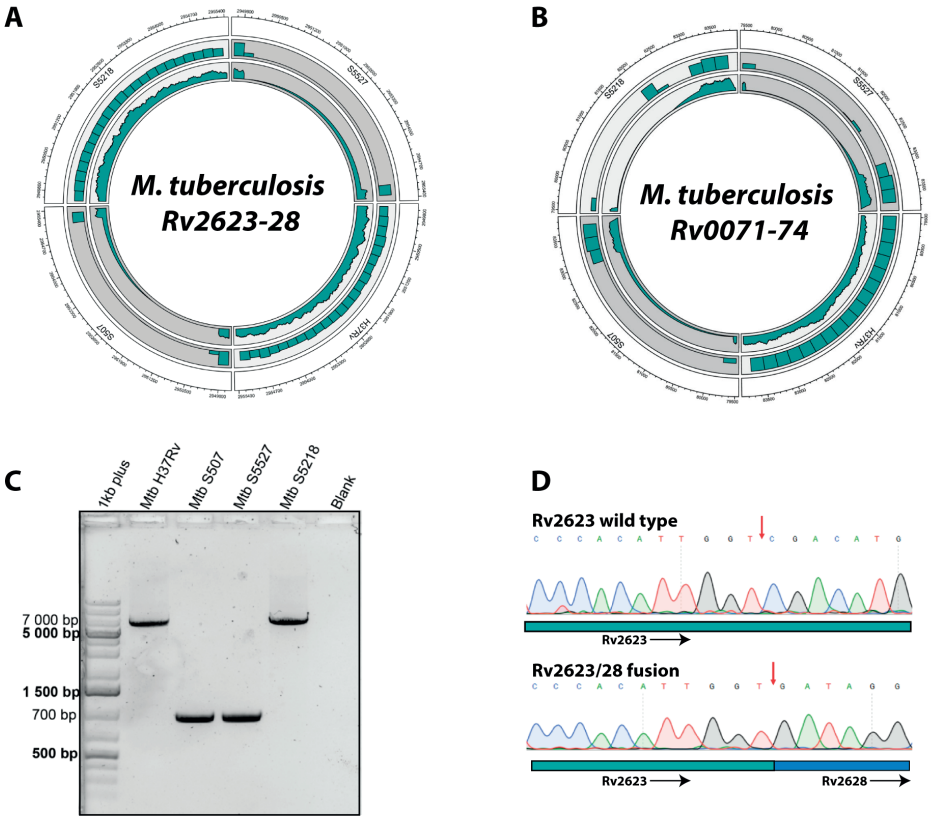



**Figure 4: *Rv2623/28* and *Rv0071/74* are gene fusions which have formed as a result of large deletions. A)** Circos plot depicting the genomic region of *Rv2623-Rv2628* (outer track) from clinical isolates S5218, S5527, S507 and H37Rv. Middle and inner tracks display read density in the region and average coverage respectively. **B)** Circos plot of *Rv0071-Rv0074* (outer track) as well as the read density (middle track) and average coverage (inner track) in the region. **C)** Polymerase chain reaction of wild type and deleted Rv2623-Rv2628 regions. **D)** Chromatograms from capillary electrophoresis displaying the deletion breakpoints (red arrow) from clinical isolate S5218 (wild type Rv2623) and S507 (Rv2623/Rv2628 fusion protein). In the circos tracks *M. tuberculosis H37Rv* is representative of the reference genotype.

MS-Digest, a tool available with the ProteinProspector software (50), was used to model the tryptic peptides of the potential chimeric proteins surrounding the expected fusion junctions of ppe38-71; Rv2623/8, Rv0071/74, and Pks15/1 as well as their wild type counterparts. Translation of the *de novo* assembly of the *ppe38/71* operon predicts the fusion junction to have the amino acid sequence S G P I A S (Additional file 2, book 1). The fusion junction amino acid sequence corresponding to V I G R is expected if the Rv2623/8 fusion is present in S507 and S5527, while D M S K is expected in wild type S5218 (Additional file 2, book 2). If *Rv0071/74* is produced in the lineage 2 strains the fusion junction V V G V G R should be detected. (Additional file 2, book 3). Lastly, if Pks15/1 is present an amino acid sequence corresponding to V P W V I S A R is expected (Additional file 2, book 4).



We used *de novo* assemblies to determine the fusion junctions of *Rv2623/28* and *Rv0071/74* genetically. The *Rv2623-Rv2628* deletion results in a new fusion gene with a restored reading frame (Fig. 5A). The *Rv0071-Rv0074* deletion has an in-frame breakpoint at position 93 (gtc) of the *Rv0071* gene, with a codon for alanine (31$^{st}$ amino acid) as well as position 289 (gtg) of *Rv0074* which codes for valine (97$^{th}$ amino acid) and thus closing the frame to create a V V G V G R amino acid sequence (Fig. 5B). This fusion candidate has recently been demonstrated to indeed encode for a functional chimeric protein, demonstrating a functional phenotype associated with this genotype (18). Next, we used our custom gene fusion database as a reference for tandem mass spectra searches to detect Rv2623/28 and Rv0071/74 in clinical isolate S507 and S5527 while using S5218 as a control. Tandem mass spectra from previously published S3651 were also searched by using the same approach to probe for *ppe38/71*. We detected the presence of a peptide that spans the Rv2623/28 fusion junction, thereby providing evidence for the expression of another chimeric protein in *M. tuberculosis* (Fig. 5C). While figure 5C represents the most confident peptide identification by the Andromeda search engine, in this specific peptide, fragmentation did not cover the fusion event located in the y1-y7 ion range. We found seven peptide spectrum matches of fragmentation across the fusion junction, thereby supporting the identification of this peptide as spanning the fusion junction between Rv2623 and Rv2628 (Additional file 1: Fig. S8 A1-A7). In addition, we were able to detect peptides corresponding to wild type Rv2623 in clinical isolate S5218 (Fig. 5D). We could not detect a peptide corresponding to the wild type Rv2623 protein in S507 or S5527, and likewise no fusion peptide was detected in S5218 (Additional file 1: Fig S8B). In addition, peptides spanning the fusion junction for Pks15/1, previously reported as an insertion was also detected in our mass spectrometry analysis (Additional file 1: Fig. S9) (22).

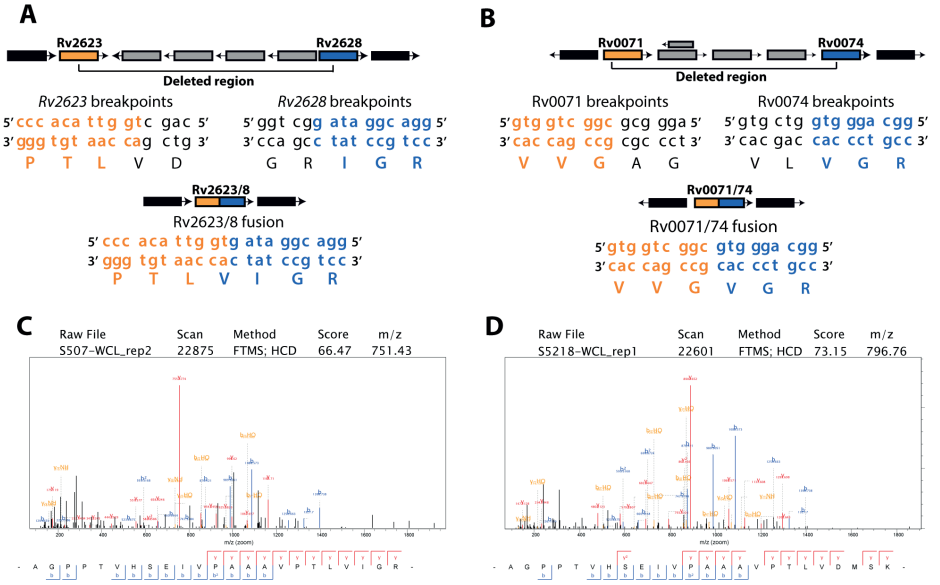



**Figure 5: Targeted *de novo* assemblies and tandem mass spectrometry identifies Rv2623/28 as chimeric protein. A)** Schematic illustration of *Rv2623-Rv2628* de novo assembly and contig ordering from clinical isolate S507. **B)** Schematic illustration of de novo assembly and contig ordering from S507 displaying the deletion region and in-frame translation of *Rv0071/74.* Black indicates out of bounds genes and grey indicates deleted genes. Tandem mass spectra of peptides representing the **C)** Rv2623/28 fusion junction from S507 and **D)** wild type Rv2623 of S5218 in the same location. False discovery rate cut-offs for assigning peptides was set at 0.01.

No peptides corresponding to PPE38/71 or Rv0071/74 fusion junctions were detected by mass spectrometry analysis. This may be due to several reasons such as low abundance, low coverage of the proteome, or in the case of Rv0071/74 a large amount of proline repeats across the fusion junction. Nevertheless Rv0071/74 was recently shown to form a chimeric protein and to be involved in the remodelling of the bacterial cell wall and have been associated with increased drug resistance (18). Furthermore, using polyclonal antiserum directed against PPE38, we were able to show that a chimeric protein was produced in strains containing the PPE38/71 fusion. Taken together, it is evident that gene fusions are not only present within *M. tuberculosis* but can also encode functional proteins.

## DISCUSSION

In this study we present a computational method to discover chimeric proteins using large scale omics data. By using unique features found in individual genomes in a proteogenomics approach, structures such as chimeric proteins can be identified in

the proteomes of clinical *M. tuberculosis* isolates. While similar methodology has been implemented in other biological fields, such as cancer research (51), studies have not yet been conducted in a high throughput fashion in bacteria. Using the same methodology proposed here, the same approach can be broadly applied to other bacteria as genomics and proteomics data becomes more prevalent.

It has been hypothesised that *M. tuberculosis* does not undergo horizontal gene transfer (11). This imposes a significant limitation on the evolutionary capabilities of the bacilli, especially when it is faced with constant evolutionary pressure due to the harsh environment of the phagosome (52). Proteins that have two distinct domains likely arose from two ancestral genes through gene fusion formation. These can be detected by comparative genomics between two or more related organisms. Gene fusion formation compress the coding potential of the genome by creating multifunctional proteins through the combination of domains (53–55). This is especially effective if there is a clear evolutionary link between multiple species from which gene fusions can be found using tools such as MosaicFinder, DomainTeam or machine learning approaches (56–58). Therefore, the formation of gene fusions could provide interesting insights into the functional evolutionary biology of *M. tuberculosis*. This is especially useful under strictive environments where reductive evolution is present (4). This is of importance as the bulk of studies mainly focus on single nucleotide variants and their functional role while larger structural genomic variation is used for the purpose of strain typing (59, 60). Interestingly, these large deletions are optimal for differentiation between lineages of *M. tuberculosis* as they are highly conserved according to a geographical origin (61). This is indicative of restrictive divergent evolution, where deletions occur frequently as a function of intracellular lifestyle yet are selected for by variable environmental conditions. The *ppe38-71* operon is a hypervariable region which arises due to the prevalence of transposon insertions (43). Therefore the deletion can occur in both lineages and not associated with a single event. However, the specific breaks in this region can have seemingly different consequences. Apart from loss of function, as a result of large-scale deletions, we demonstrate that these deletions can result in the formation of detectable chimeric proteins, which in turn likely has functional consequences. We could demonstrate such functionality for the *ppe38/71* fusion protein in the form of PE-PGRS secretion. Previous reports have shed light on the increased ability of lineage 2 *M. tuberculosis* to transmit and cause disease (62, 63). As the lack of *ppe38-71* is also associated with increased virulence (13), it is likely compounding and contributing to the increased disease-causing capability observed in lineage 2 isolates (13). Furthermore, others have demonstrated functionality for Rv0071/74 (18), Rv3346c/55c (45) and PKS15/1 (22) chimeric proteins as well as the impact of large deletions on increased virulence (64). *M. tuberculosis* strains can be grouped into sub-lineages by occurrence

of large deletions which remove multiple genes in the process. This has the clear effect of abolishing the functions of the genes that have been lost, but also as seen in the case of RD107 and $RD^{RIO}$ can form new coding sequences in the process. Contrary to the *ppe38-71* deletions, formation of the gene fusions are likely a result of selection and expanded to create sub-lineages and thus form a monophyletic group.

A new chimeric protein candidate, Rv2623/28, was identified in this study. The parent proteins for this chimera have been associated with entrance into dormancy (65, 66) and associated with latency (67). Overexpression of Rv2623 causes a decrease in proliferation of *M. tuberculosis in vitro* and increased pathology in mice thereby mediating entrance into dormancy (66). This protein is comprised of two ATP binding domains and the chimera Rv2623/28 has one of these domains followed by a Rv2628 N-terminus. If this domain remains stabilised with the Rv2628 N-terminus, it could result in a less pronounced inhibition of growth than reported for the full length Rv2623 (66). Furthermore, the gene at the 3' end of the deletion, *Rv2628* is associated with latent tuberculosis infection as shown by a stronger cumulative IFN-gamma response towards Rv2628 antigens, compared to tuberculosis positive individuals (68, 69). The *Rv2623-Rv2628* gene fusion is therefore an interesting open reading frame potentially with aspects associated with both early and late stage dormancy.

Information on transcribed gene fusions can also be identified by RNA-sequencing by using split read transcripts with multiple tools available for this purpose (70–72). A powerful approach presented by RNA sequencing is the use of *de novo* transcript assembly to mitigate the requirement of the reference sample, however this suffers a penalty in the form of decreased accuracy (73). While RNA-sequencing provides a powerful tool to identify gene fusions, the function of the resulting chimera is performed on the protein level. This is especially important as there have been reports on the poor correlation between RNA and protein content (74, 75). This general discrepancy exists due to the complex mechanisms governing mRNA and protein regulation, both post-transcriptionally as well as post-translationally and likely has a significant temporal component (76). Therefore, the combination of genomics and proteomics provides a powerful, yet under developed, approach to high throughput identification of expressed gene fusions. The combination of these technologies does however bring with it limitations associated with each respective platform. These limitations influence the detection power of a combined approach. From a genomics perspective, detection of gene fusions is influenced by the choice of a reference strain, this could however be mitigated by using a known ancestor or a metagenome approach by combining various related genomes in order to cover a broad range of genes (77). Long read sequencing provides an alternative option to facilitate the process of resolving gene fusions in the

genome. This can be used as either a *de novo* assembly, reference based assembly or a hybrid approach to genetically detect gene fusions in a similar fashion as the short reads used here. With longer reads the exact breakpoints could be resolved with increased accuracy and thus result in the detection of more gene fusions. This is especially valuable as failure of alignments to pinpoint exact breakpoints can confound detection and either falsely call a gene fusion or miss potential gene fusions completely. Finally, a diverse library of sequences is necessary in order to effectively search for gene fusions using whole genome sequencing. In the Western Cape region of South Africa, where this study was conducted, the majority of *M. tuberculosis* strains are either members of lineage 2 or lineage 4. We therefore focussed on these lineages of *M. tuberculosis* to act as our library and *M. tuberculosis* H37Rv as the reference, thus limiting the amount of gene fusions we could detect. As there are other lineages, characterised by large sequence polymorphisms, it is likely that there are more gene fusions to be identified.



The use of data-dependent tandem mass spectrometry analysis also limits the gene fusion detection ability. As the mass spectrometer only selects the "topN" most intense precursor ions (78), a gene fusion would need to have a relatively high abundance to guarantee detection. In addition, proteins are not completely sequenced using shotgun proteomics, thus fusion peptides that do not contain trypsin cleavage sites will be missed by this approach (79). As the fusion junction, and thus the fusion protein, is resolved by a single peptide the likelihood for detection decreases as well. With arguably low genomic diversity in this cohort, where only two major lineages were screened, we still identified four candidate gene fusions using a genomics and proteomics approach. *M. tuberculosis* is subject to reductive evolution, which is a marked characteristic of an intracellular lifestyle (1, 80, 81). With this limited and decreasing coding potential of *M. tuberculosis*, even rare occurrences of gene fusions and resulting chimeric proteins could result in significant phenotypic consequences. Some of the limitations to detect chimeric proteins in *M. tuberculosis* proteomes can however be overcome using a targeted proteomics approach to identify fusion junction peptides or using dedicated spectral libraries in conjunction with large scale data independent approaches such as SWATH-MS (82, 83) .

In this study, we focussed on gene fusions and their detection. However, the use of similar methodology can be extended to other phenotypic features typically lost by reference-based assembly, such as identifying novel proteins from unmapped reads (19). Indeed, a significant number of tandem mass spectra remain unassigned after database searches (84), of which many display high quality spectra (85). Spectra remain unassigned due to a multitude of reasons such as charge state, fully tryptic searches, post-translational or chemical modifications and errors in mass to charge

measurements (50, 86). It is also reasonable to assume that incomplete databases and unrepresented proteins also contribute to the large number of unassigned spectra. By expanding methodologies presented here and with further exploration of cross-platform omics technologies, previously hidden features can be extracted and provide a high throughput approach to identifying novel features. In this study *M. tuberculosis* was used, however the same methodology can be implemented for other bacteria as well.

In conclusion, here we demonstrated that additional and often overlooked features of the genome contribute to the physiology of *M. tuberculosis*. These features provide a means to utilise vertical evolution to gain functionality in the absence of horizontal gene transfer. Further study into the effects of structural variation along with the mechanism that allow these features to translate to functional phenotypes could be important to understanding *M. tuberculosis* disease-causing capabilities.

## AVAILABILITY

The gene fusion calling software is freely available under the GNU general public licence version 3 available (https://github.com/JamesGallant/Genomics). A Docker image of the software is also available at the following URL: https://hub.docker.com/r/jamesgallant/pegasus

Accession numbers for genomes are available in supplementary table 1 and all the raw genomics data has been deposited in the European Nucleotide Archive (ENA) under the project accession PRJEB36366

Accession number for proteomes and custom FASTA file available on ProteomeXchange: PXD017298

## FUNDING

This work was supported by the National Research Foundation/Vrije Universiteit Desmond Tutu Doctoral Training Program awarded to JG. Research done in the laboratory of SLS was supported financially by the SA MRC Centre for TB Research and DST/NRF Centre of Excellence for Biomedical Tuberculosis Research. SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539.

Research in AP laboratory was funded by a faculty baseline grant (BAS/1/1020-01-01) from KAUST.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

**JLG**: Conceptualisation, methodology, software, validation, formal analysis, investigation, data curation, visualisation, writing

**JM**: validation, investigation, visualisation, writing

**RU**: investigation, visualisation

**CTHJ**: investigation, visualisation

**NK**: investigation, validation

**AP**: investigation, resources

**RW**: resources, funding acquistion

**WB**: Conceptualisation, resources, project administration, funding acquisition, supervision, writing

**TH**: Conceptualisation, methodology, formal analysis, funding acquisition, supervision, writing

**SS**: Conceptualisation, resources, project administration, funding acquisition, supervision, writing

## SUPPLEMENTARY DATA

Supplementary data for not in this document can be accessed at the following URL with a Mendeley account:

**https://tinyurl.com/5u3sumvm**

## REFERENCES


1. Weinert,L.A. and Welch,J.J. (2017) Why Might Bacterial Pathogens Have Small Genomes? *Trends Ecol. Evol.*, **32**, 936–947.
2. Tamas,I. (2002) 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. *Science (80-. ).*, **296**, 2376–2379.
3. Vissa,V.D. and Brennan,P.J. (2001) The genome of Mycobacterium leprae: a minimal mycobacterial gene set. *Genome Biol.*, **2**, REVIEWS1023.
4. Dé,F., Veyrier,R.J., Dufort,A. and Behr,M.A. (2011) The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol.*, **19**, 156–161.
5. Flores,L., Van,T., Narayanan,S., DeRiemer,K., Kato-Maeda,M. and Gagneux,S. (2007) Large sequence polymorphisms classify Mycobacterium tuberculosis strains with ancestral spoligotyping patterns. *J. Clin. Microbiol.*, **45**, 3393–5.
6. Hall,J.P.J., Brockhurst,M.A. and Harrison,E. (2017) Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philos. Trans. R. Soc. B Biol. Sci.*, **372**, 20160424.
7. Fournier,G.P. and Gogarten,J.P. (2008) Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. *J. Bacteriol.*, **190**, 1124–7.
8. Boto,L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B Biol. Sci.*, **277**, 819–827.
9. Popa,O., Hazkani-Covo,E., Landan,G., Martin,W. and Dagan,T. (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.*, **21**, 599–609.
10. Philippot,L., Andersson,S.G.E., Battin,T.J., Prosser,J.I., Schimel,J.P., Whitman,W.B. and Hallin,S. (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.*, **8**, 523–529.
11. Namouchi,A., Didelot,X., Schock,U., Gicquel,B. and Rocha,E.P.C. (2012) After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Res.*, **22**, 721–734.
12. Stucki,D. and Gagneux,S. (2013) Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis*, **93**, 30–39.
13. Ates,L.S., Dippenaar,A., Ummels,R., Piersma,S.R., van der Woude,A.D., van der Kuij,K., Le Chevalier,F., Mata-Espinosa,D., Barrios-Payán,J., Marquina-Castillo,B., *et al.* (2018) Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nat. Microbiol.*, **3**, 181–188.
14. Xia,L.C., Bell,J.M., Wood-Bouwens,C., Chen,J.J., Zhang,N.R. and Ji,H.P. (2018) Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.*, **46**, e19.
15. Li,H., Wang,J., Ma,X. and Sklar,J. (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle*, **8**, 218–222.
16. Alfaro,J.A., Sinha,A., Kislinger,T. and Boutros,P.C. (2014) Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nat. Methods*, **11**, 1107–1113.
17. Farr,A.D., Remigi,P. and Rainey,P.B. (2017) Adaptive evolution by spontaneous domain fusion and protein relocalization. *Nat. Ecol. Evol.*, **1**, 1562–1568.

18. Qin,L., Wang,J., Lu,J., Yang,H., Zheng,R., Liu,Z., Huang,X., Feng,Y., Hu,Z. and Ge,B. (2019) A deletion in the RD105 region confers resistance to multiple drugs in Mycobacterium tuberculosis. *BMC Biol.*, **17**, 7.
19. Heunis,T., Dippenaar,A., Warren,R.M., van Helden,P.D., van der Merwe,R.G., Gey van Pittius,N.C., Pain,A., Sampson,S.L. and Tabb,D.L. (2017) Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium tuberculosis* Isolates. *J. Proteome Res.*, **16**, 3841–3851.
20. Warren,R.M., Richardson,M., Sampson,S.L., van der Spuy,G.D., Bourn,W., Hauman,J.H., Heersma,H., Hide,W., Beyers,N. and van Helden,P.D. (2001) Molecular evolution of Mycobacterium tuberculosis: phylogenetic reconstruction of clonal expansion. *Tuberculosis*, **81**, 291–302.
21. Groenen,P.M.A., Bunschoten,A.E., Soolingen,D. van and Errtbden,J.D.A. van (1993) Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol. Microbiol.*, **10**, 1057–1065.
22. Chaiprasert,A., Yorsangsukkamol,J., Prammananan,T., Palittapongarnpim,P., Leechawengwong,M. and Dhiraputra,C. (2006) Intact pks15/1 in non-W-Beijing Mycobacterium tuberculosis isolates. *Emerg. Infect. Dis.*, **12**, 772–4.
23. Phelan,J.E., O'Sullivan,D.M., Machado,D., Ramos,J., Oppong,Y.E.A., Campino,S., O'Grady,J., McNerney,R., Hibberd,M.L., Viveiros,M., *et al.* (2019) Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.*, **11**, 41.
24. Warren,R., de Kock,M., Engelke,E., Myburgh,R., Gey van Pittius,N., Victor,T. and van Helden,P. (2006) Safe Mycobacterium tuberculosis DNA Extraction Method That Does Not Compromise Integrity. *J. Clin. Microbiol.*, **44**, 254–256.
25. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–303.
26. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jage,B.B., Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S. V., Eiglmeier,K., Gas,S., *et al.* (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.
27. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
28. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
29. Rausch,T., Zichner,T., Schlattl,A., Stütz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
30. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
31. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

32. Chikhi,R. and Medvedev,P. (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, **30**, 31–37.
33. Luo,R., Liu,B., Xie,Y., Li,Z., Huang,W., Yuan,J., He,G., Chen,Y., Pan,Q., Liu,Y., *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
34. Assefa,S., Keane,T.M., Otto,T.D., Newbold,C. and Berriman,M. (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, **25**, 1968–9.
35. Rombel,I.T., Sykes,K.F., Rayner,S. and Johnston,S.A. (2002) ORF-FINDER: a vector for high-throughput gene identification. *Gene*, **282**, 33–41.
36. Ramagli,L.S. and Rodriguez,L. V. (1985) Quantitation of microgram amounts of protein in two-dimensional polyacrylamide gel electrophoresis sample buffer. *Electrophoresis*, **6**, 559–563.
37. Juri Rappsilber,†, Yasushi Ishihama,†,‡ and and Mann*,M. (2002) Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. 10.1021/AC026117I.
38. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J. V. and Mann,M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
39. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
40. Jong,W.S.P., Vikström,D., Houben,D., van den Berg van Saparoea,H.B., de Gier,J.-W. and Luirink,J. (2017) Application of an E. coli signal sequence as a versatile inclusion body tag. *Microb. Cell Fact.*, **16**, 50.
41. Abdallah,A.M., Verboom,T., Weerdenburg,E.M., Gey van Pittius,N.C., Mahasha,P.W., Jiménez,C., Parra,M., Cadieux,N., Brennan,M.J., Appelmelk,B.J., *et al.* (2009) PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. *Mol. Microbiol.*, **73**, 329–340.
42. Harboe,M., Malin,A.S., Dockrell,H.S., Wiker,H.G., Ulvund,G., Holm,A., Jørgensen,M.C. and Andersen,P. (1998) B-cell epitopes and quantification of the ESAT-6 protein of Mycobacterium tuberculosis. *Infect. Immun.*, **66**, 717–23.
43. McEvoy,C.R.E., van Helden,P.D., Warren,R.M., van Pittius,N. and Gey van Pittius,N.C. (2009) Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. *BMC Evol. Biol.*, **9**, 237.
44. Burggraaf,M.J., Speer,A., Meijers,A.S., Ummels,R., Van Der Sar,A.M., Korotkov,K. V., Bitter,W. and Kuijl,C. (2019) Type VII secretion substrates of pathogenic mycobacteria are processed by a surface protease. *MBio*, **10**.
45. Lazzarini,L.C.O., Huard,R.C., Boechat,N.L., Gomes,H.M., Oelemann,M.C., Kurepina,N., Shashkina,E., Mello,F.C.Q., Gibson,A.L., Virginio,M.J., *et al.* (2007) Discovery of a novel Mycobacterium tuberculosis lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J. Clin. Microbiol.*, **45**, 3891–3902.
46. Tsolaki,A.G., Gagneux,S., Pym,A.S., de la Salmoniere,Y.O., Kreiswirth,B.N., Van,S.D. and Small,P.M. (2005) Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. *J Clin Microbiol*, **43**.

47. Målen,H., Pathak,S., Søfteland,T., de Souza,G.A. and Wiker,H.G. (2010) Definition of novel cell envelope associated proteins in Triton X-114 extracts of Mycobacterium tuberculosis H37Rv. *BMC Microbiol.*, **10**, 132.
48. Liu,Y., Harrison,P.M., Kunin,V. and Gerstein,M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, **5**, R64.
49. Tsolaki,A.G., Gagneux,S., Pym,A.S., Goguet de la Salmoniere,Y.-O.L., Kreiswirth,B.N., Van Soolingen,D. and Small,P.M. (2005) Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. *J. Clin. Microbiol.*, **43**, 3185–91.
50. Chalkley,R.J., Baker,P.R., Huang,L., Hansen,K.C., Allen,N.P., Rexach,M. and Burlingame,A.L. (2005) Comprehensive Analysis of a Multidimensional Liquid Chromatography Mass Spectrometry Dataset Acquired on a Quadrupole Selecting, Quadrupole Collision Cell, Time-of-flight Mass Spectrometer. *Mol. Cell. Proteomics*, **4**, 1194–1204.
51. Sun,H., Xing,X., Li,J., Zhou,F., Chen,Y., He,Y., Li,W., Wei,G., Chang,X., Jia,J., *et al.* (2013) Identification of gene fusions from human lung cancer mass spectrometry data. *BMC Genomics*, **14 Suppl 8**, S5.
52. Ehrt,S. and Schnappinger,D. (2009) Mycobacterial survival strategies in the phagosome: defence against host stresses. *Cell. Microbiol.*, **11**, 1170–8.
53. Bornberg-Bauer,E., Beaussart,F., Kummerfeld,S.K., Teichmann,S.A. and Weiner,J. (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.*, **62**, 435–445.
54. Orengo,C.A. and Thornton,J.M. (2005) PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE. *Annu. Rev. Biochem.*, **74**, 867–900.
55. Pasek;,S., Risler,J.-L. and Brézellec,P. (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **22**, 1418–1423.
56. MosaicFinder: identification of fused gene families in sequence similarity networks | Bioinformatics | Oxford Academic.
57. Pasek,S., Bergeron,A., Risler,J.L., Louis,A., Ollivier,E. and Raffinot,M. (2005) Identification of genomic features using microsyntenies of domains: Domain teams. *Genome Res.*, **15**, 867–874.
58. Henry,C.S., Lerma-Ortiz,C., Gerdes,S.Y., Mullen,J.D., Colasanti,R., Zhukov,A., Frelin,O., Thiaville,J.J., Zallot,R., Niehaus,T.D., *et al.* (2016) Systematic identification and analysis of frequent gene fusion events in metabolic pathways. *BMC Genomics*, **17**, 473.
59. Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.-S., Day,N.P.J., Enright,M.C., Goldstein,R., Hood,D.W., Kalia,A., Moore,C.E., *et al.* (2001) Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci.*, **98**, 182–187.
60. Ho,T.B.L., Robertson,B.D., Taylor,G.M., Shaw,R.J. and Young,D.B. (2000) Comparison of *Mycobacterium tuberculosis* Genomes Reveals Frequent Deletions in a 20 kb Variable Region in Clinical Isolates. *Yeast*, **1**, 272–282.
61. Reed,M.B., Pichler,V.K., McIntosh,F., Mattia,A., Fallow,A., Masala,S., Domenech,P., Zwerling,A., Thibert,L., Menzies,D., *et al.* (2009) Major Mycobacterium tuberculosis lineages associate with patient country of origin. *J. Clin. Microbiol.*, **47**, 1119–28.
62. de Jong,B.C., Hill,P.C., Aiken,A., Awine,T., Antonio,M., Adetifa,I.M., Jackson-Sillah,D.J., Fox,A., DeRiemer,K., Gagneux,S., *et al.* (2008) Progression to Active Tuberculosis, but Not

Transmission, Varies by *Mycobacterium tuberculosis* Lineage in The Gambia. *J. Infect. Dis.*, **198**, 1037–1043.

63. Hanekom,M., Spuy,G.D. v. a. n. d. e. r., Streicher,E., Ndabambi,S.L., McEvoy,C.R., Kidd,M., Beyers,N., Victor,T.C., van Helden,P.D. and Warren,R.M. (2007) A recently evolved sublineage of the Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease. *J Clin Microbiol*, **45**.
64. ten Bokum,A.M.C., Movahedzadeh,F., Frita,R., Bancroft,G.J. and Stoker,N.G. (2008) The case for hypervirulence through gene deletion in Mycobacterium tuberculosis. *Trends Microbiol*, **16**, 436–441.
65. Butcher,P.D., Betts,J., Banerjee,D.K. and Monahan,I.M. (2001) Differential expression of mycobacterial proteins following phagocytosis by macrophages. *Microbiology*, **147**, 459–471.
66. Drumm,J.E., Mi,K., Bilder,P., Sun,M., Lim,J., Bielefeldt-Ohmann,H., Basaraba,R., So,M., Zhu,G., Tufariello,J.M., *et al.* (2009) Mycobacterium tuberculosis Universal Stress Protein Rv2623 Regulates Bacillary Growth by ATP-Binding: Requirement for Establishing Chronic Persistent Infection. *PLoS Pathog.*, **5**, e1000460.
67. Goletti,D., Butera,O., Vanini,V., Lauria,F.N., Lange,C., Franken,K.L.M.C., Angeletti,C., Ottenhoff,T.H.M. and Girardi,E. (2010) Response to Rv2628 latency antigen associates with cured tuberculosis and remote infection. *Eur. Respir. J.*, **36**, 135–42.
68. Leyten,E.M.S., Lin,M.Y., Franken,K.L.M.C., Friggen,A.H., Prins,C., van Meijgaarden,K.E., Voskuil,M.I., Weldingh,K., Andersen,P., Schoolnik,G.K., *et al.* (2006) Human T-cell responses to 25 novel antigens encoded by genes of the dormancy regulon of Mycobacterium tuberculosis. *Microbes Infect.*, **8**, 2052–2060.
69. Black,G.F., Thiel,B.A., Ota,M.O., Parida,S.K., Adegbola,R., Boom,W.H., Dockrell,H.M., Franken,K.L.M.C., Friggen,A.H., Hill,P.C., *et al.* (2009) Immunogenicity of novel DosR regulon-encoded candidate antigens of Mycobacterium tuberculosis in three high-burden populations in Africa. *Clin. Vaccine Immunol.*, **16**, 1203–12.
70. Iyer,M.K., Chinnaiyan,A.M. and Maher,C.A. (2011) ChimeraScan: A tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903–2904.
71. Li,Y., Chien,J., Smith,D.I. and Ma,J. (2011) FusionHunter: Identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, **27**, 1708–1710.
72. Jia,W., Qiu,K., He,M., Song,P., Zhou,Q., Zhou,F., Yu,Y., Zhu,D., Nickerson,M.L., Wan,S., *et al.* (2013) SOAPfuse: An algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.*, **14**.
73. Haas,B.J., Dobin,A., Li,B., Stransky,N., Pochet,N. and Regev,A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**.
74. Gry,M., Rimini,R., Strömberg,S., Asplund,A., Pontén,F., Uhlén,M. and Nilsson,P. (2009) Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, **10**, 365.
75. Cortes,T., Schubert,O.T., Banaei-Esfahani,A., Collins,B.C., Aebersold,R. and Young,D.B. (2017) Delayed effects of transcriptional responses in Mycobacterium tuberculosis exposed to nitric oxide suggest other mechanisms involved in survival. *Sci. Rep.*, **7**.
76. Liu,Y., Beyer,A. and Aebersold,R. (2016) Leading Edge Review On the Dependency of Cellular Protein Levels on mRNA Abundance. 10.1016/j.cell.2016.03.014.

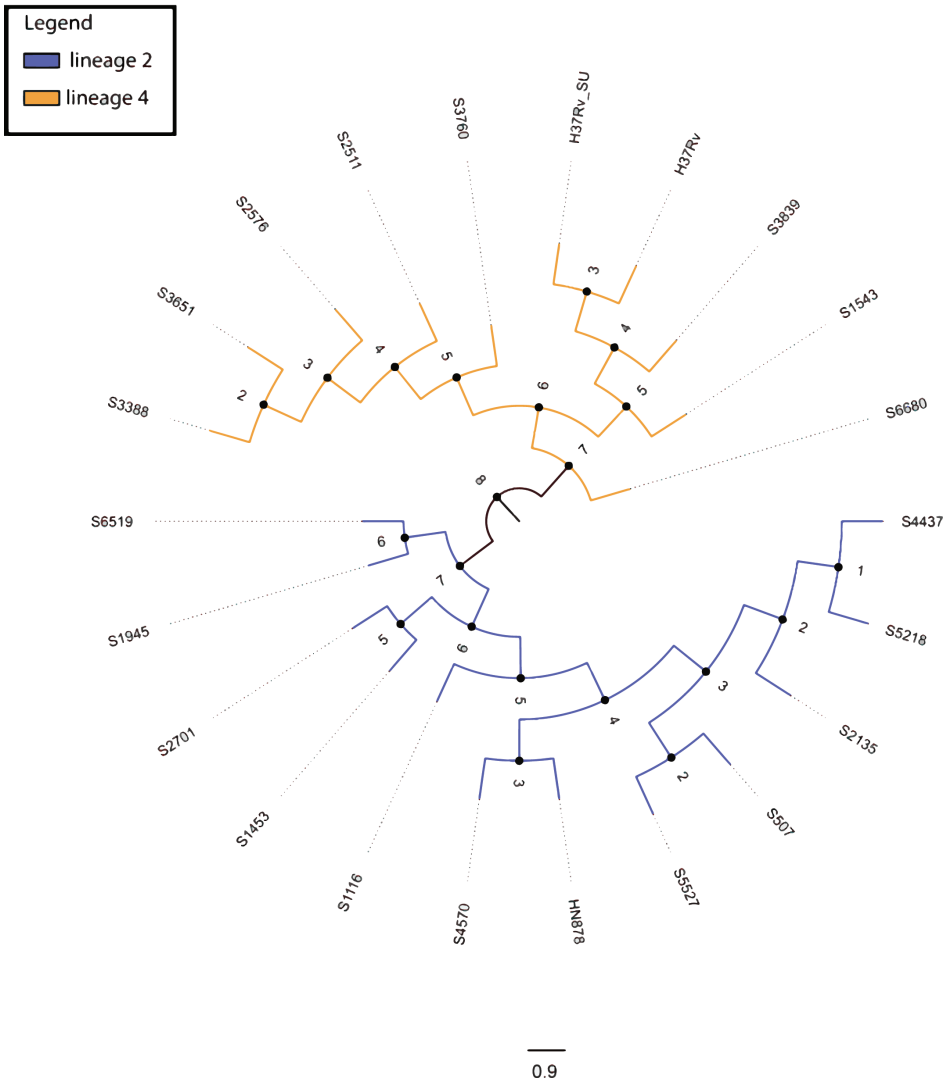
77. Riesenfeld,C.S., Schloss,P.D. and Handelsman,J. (2004) Metagenomics: Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.*, **38**, 525–552.
78. Wang,N. and Li,L. (2008) Exploring the Precursor Ion Exclusion Feature of Liquid Chromatography–Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry for Improving Protein Identification in Shotgun Proteome Analysis. *Anal. Chem.*, **80**, 4696–4710.
79. Raulfs,M.D.M., Breci,L., Bernier,M., Hamdy,O.M., Janiga,A., Wysocki,V. and Poutsma,J.C. (2014) Investigations of the Mechanism of the "Proline Effect" in Tandem Mass Spectrometry Experiments: The "Pipecolic Acid Effect". *J. Am. Soc. Mass Spectrom.*, **25**, 1705–1715.
80. Koskiniemi,S., Sun,S., Berg,O.G. and Andersson,D.I. (2012) Selection-Driven Gene Loss in Bacteria. *PLoS Genet*, **8**, e1002787.
81. Zimpel,C.K., Brandão,P.E., de Souza Filho,A.F., de Souza,R.F., Ikuta,C.Y., Neto,J.S.F., Soler Camargo,N.C., Heinemann,M.B. and Guimarães,A.M.S. (2017) Complete genome sequencing of Mycobacterium bovis SP38 and comparative genomics of Mycobacterium bovis and M. tuberculosis strains. *Front. Microbiol.*, **8**.
82. Gillet,L.C., Navarro,P., Tate,S., Röst,H., Selevsek,N., Reiter,L., Bonner,R. and Aebersold,R. (2012) Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics*, **11**, O111.016717.
83. Rauniyar,N. (2015) Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. *Int. J. Mol. Sci.*, **16**, 28566–28581.
84. Nesvizhskii,A.I., Roos,F.F., Grossmann,J., Vogelzang,M., Eddes,J.S., Gruissem,W., Baginsky,S. and Aebersold,R. (2006) Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data. *Mol. Cell. Proteomics*, **5**, 652–670.
85. Flikka,K., Martens,L., Vandekerckhove,J., Gevaert,K. and Eidhammer,I. (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, **6**, 2086–2094.
86. Nielsen,M.L., Savitski,M.M. and Zubarev,R.A. (2006) Extent of Modifications in Human Proteome Samples and Their Effect on Dynamic Range of Analysis in Shotgun Proteomics. *Mol. Cell. Proteomics*, **5**, 2384–2391.

## SUPPLEMENTARY FIGURES

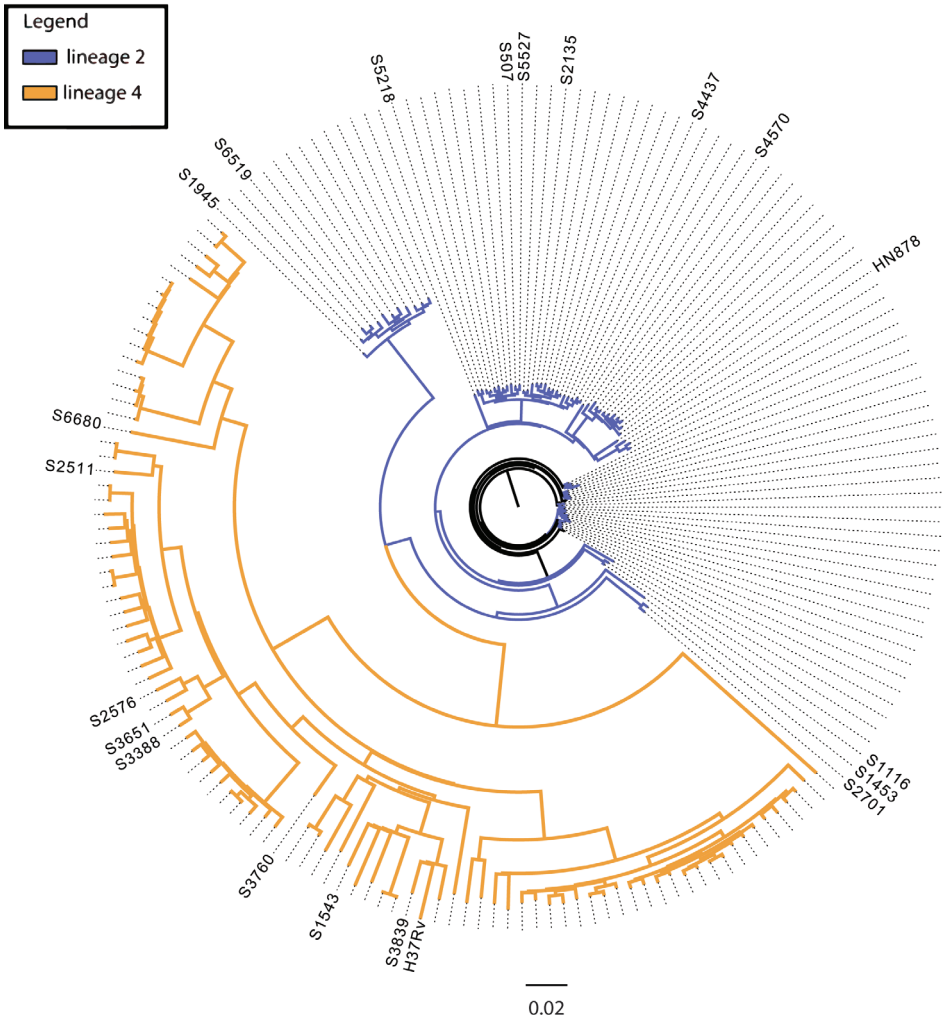
**Supplementary table 1: *Mycobacterium tuberculosis* lineage and typing/**

| Sample ID | IS6110_3 | Spoligotype pattern | International type | Family | DST results | Tbprofiler | Accession | |
|---|---|---|---|---|---|---|---|---|
| **S1116** | 484 | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1.1 | ERS4262366 |
| **S0507** | | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1 | ERS4262367 |
| **S1453** | 907 | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1 | ERS4262368 |
| **S1543** | 93 | | ND | T1 | Susc to INH/RIF | | 4,8 | ERS4262369 |
| **S2511** | 536 | ggcgggggggggggggggggcccgggggggggccccgggggggg | 20 | LAM1 | Susc to INH/RIF | | 4.3.4.1 | ERS4262383 |
| **S2576** | 363 | gggggggggggggggggggcccgggggggggccccgggcgggg | 60 | LAM4 | Susc to INH/RIF | | 4.3.3 | ERS4262377 |
| **S3388** | 456 | ggggggccccgggggccccccgccgggggggccccccgggggggg | 2018 | Family 14 | Susc to INH/RIF | | 4.3.3 | ERS4262371 |
| **S3651** | 472 | gggggggggggggccccccccgggggggggccccgggggggg | 1737 | T1 | Susc to INH/RIF | | 4.3.3 | ERS4262372 |
| **S3760** | 700 | gggggggggggggggggggcccgggggggggccccgggcgggg | 60 | LAM4 | Susc to INH/RIF | | 4 | ERS4262376 |
| **S3839** | 512 | gggggggggggggggggggggggggggggggcccgggggggg | 53 | T1 | Susc to INH/RIF | | 4,9 | ERS4262378 |
| **S4437** | | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1.1 | ERS4262374 |
| **S4570** | 810 | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1 | ERS4262375 |
| **S5218** | | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1.1 | ERS4262379 |
| **S5527** | | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | Susc to INH/RIF | | 2.2.1.1 | ERS4262380 |
| **S6519** | 464 | ccccccccccccccccccccccccccccccccgggggggggg | 1 | BEIJING | MDR (Resistant to Kan) | | 2.2.2 | ERS4262381 |
| **S6680** | 1380 | gggggggggggggcgggggggggggggggggcgccccgggggggg | 36 | H3-T3 | Susc to INH/RIF | | 4.1.2.1 | ERS4262382 |
| **H37Rv_SU** | | gggggggggggggggggggccggggcgggggggccccgggggggg | not in SITVIT | H37Rv | Susc to INH/RIF | | | PRJEB32340 |
| **HN878** | N/A | N/A | N/A | N/A | N/A | | 2.2.1.2 | |
| **S1945** | N/A | N/A | N/A | N/A | N/A | | 2.2.2 | ERS1203569 |
| **S2135** | N/A | N/A | N/A | N/A | N/A | | 2.2.1.1 | ERS1203576 |
| **S2701** | N/A | N/A | N/A | N/A | N/A | | 2.2.1 | ERS1203570 |

Supplementary figure 1 A

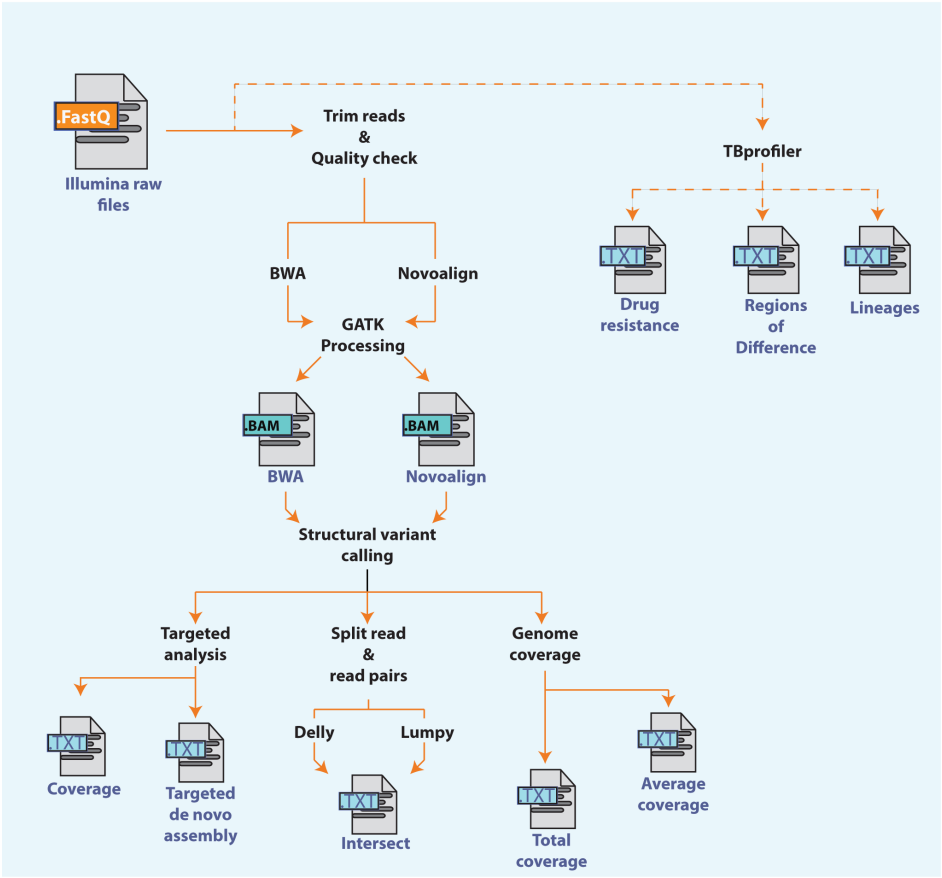


## Supplementary figure 1B

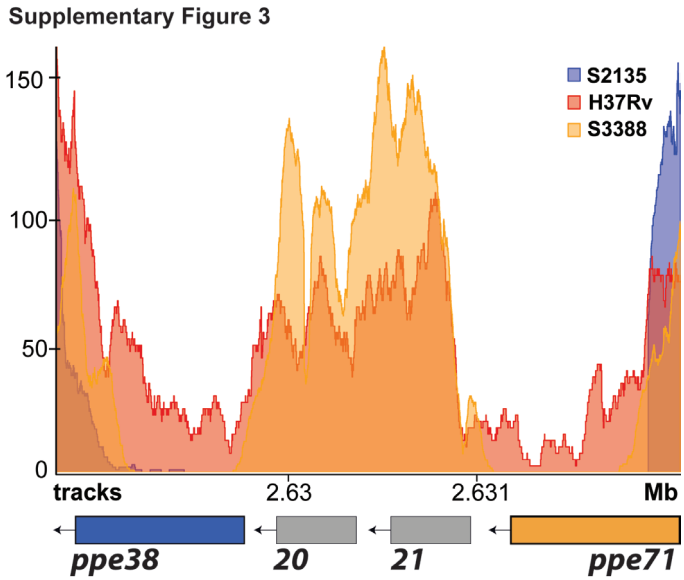



**Supplementary figure 1: Phylogenetic tree of *M. tuberculosis* clinical isolates. A)** Phylogenetic tree represents the clinical isolates typed in supplementary table 1 and **B)** represents all strains used in this study. The tree was constructed by aligning raw sequence data to *M. tuberculosis* H37Rv and identify single nucleotide variants from each isolate. The general time reversal model of nucleotide substitution was used to construct an axelerated maximum likelihood phylogeny of the isolates with 1000 bootstrap pseudo replicates. Positions with gaps or missing replicates were not considered in this analysis.

**Supplementary figure 2**

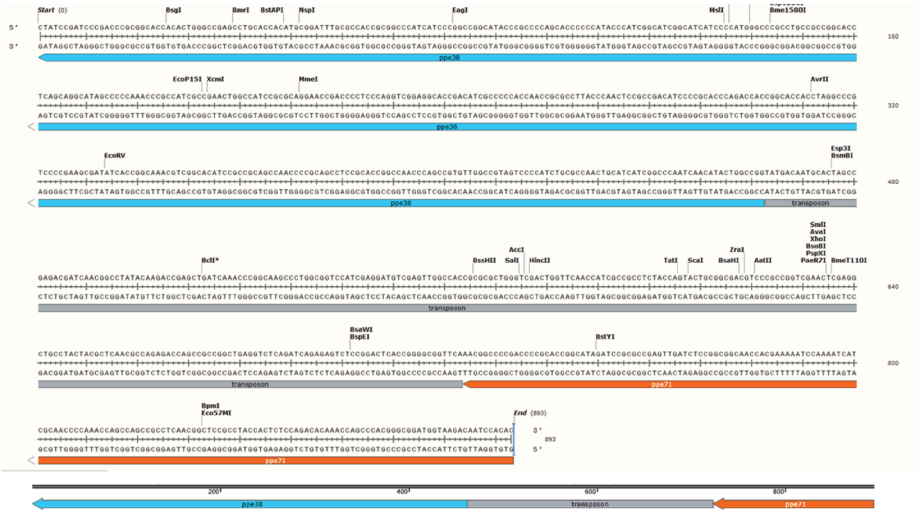



**Supplementary figure 2 Schematic diagram depicting the pipeline used to analyse *M. tuberculosis* genomes.** This software was constructed in Linux Ubuntu distribution using the Bourne again shell scripting language. Pre-processing and processing of Illumina raw files followed the Genome Analysis Toolkit best practises guidelines defined by the BROAD institute (https://software.broadinstitute.org/gatk/best-practices/workflow). Identification of deletions utilised both coverage based approaches as well as split read and read pair methods. Previously published software TBprofiler (dashed line) was used to determine *M. tuberculosis* lineages computationally and not created in this work. This software is available on GitHub (https://github.com/HostPathogenSU/Pegasus).

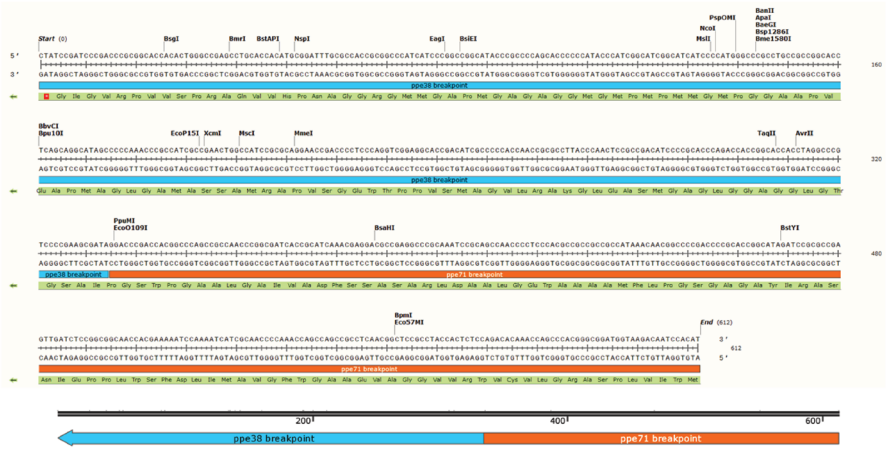



**Supplementary figure 3: Reads in the *ppe38-71* region depicting three alignment profiles observed in this study.** Each of the depicted strains display an alignment indicating a deletion (S2135), a wild type operon where reads fail, likely due to transposon insertion (S3388) and an example where reads are able to map across the operon (H37Rv). The presence of reads in *mt2420* (20) and *mt2421* (21) was considered a wild type operon which is able to secrete PE-PGRS proteins.

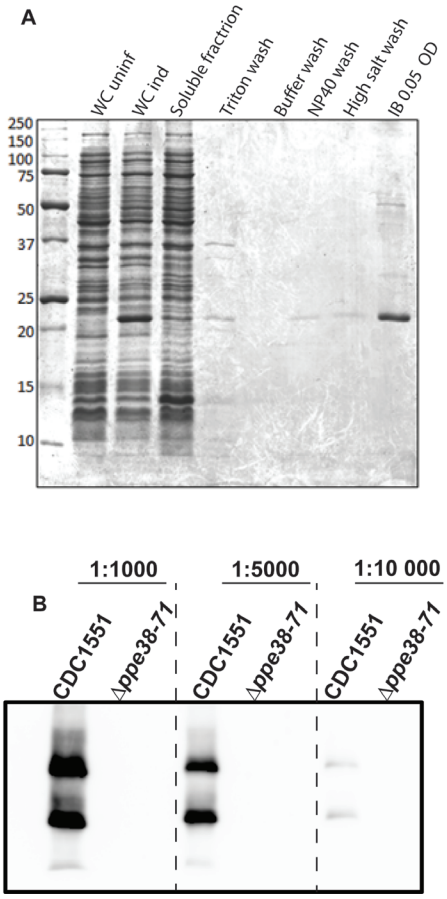
**Supplementary Figure 5A**




**Supplementary Figure 5B**

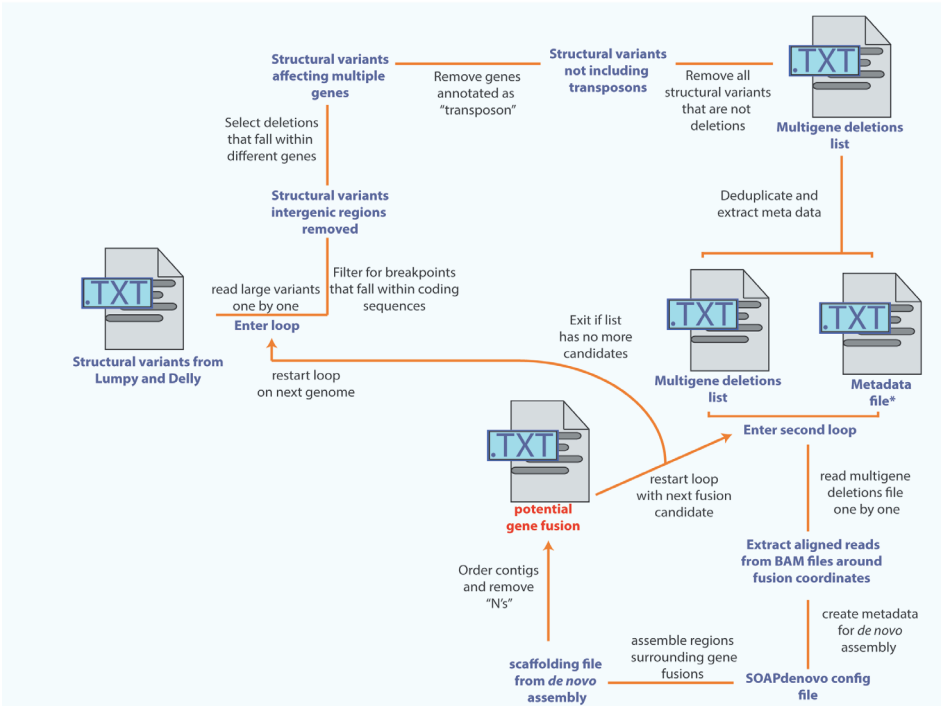



**Supplementary figure 5: Schematic representation of the *de novo* assembly from S507 and S3651 in the ppe38-71 region.** *De novo* assembly was done by extracting paired end reads flanking the breakpoints of the *ppe38-71* deletion on the 3' and 5' side from novoaligned BAM files from **A)** S507 and **B)** S3651 using Samtools. SOAPdenovo2 and abacas was used to perform the assembly and re-constitute the reads into a consensus sequence. Snapgene was used to visualise the assembly and annotations were based on cross comparison with the sequence known sequence information for *ppe38* and *ppe71*. The unknown sequence separating *ppe38* and *ppe71* in the S507 was identified as a transposon sequence by BLASTx.




**Supplementary figure 6: Isolation of ppe38 recombinant protein and testing of anti-ppe38 antiserum. A)** SDS-PAGE and coomassie stain of inclusion bodies containing PPE38 protein isolated from *E. coli* top10F' cells. This inclusion body extract was used for immunisation in rabbits. **B)** Western blot depicting anti-PPE38 rabbit antiserum at various dilutions. *M. tuberculosis* CDC1551 and *Δppe38-71* supernatant was for as the target for western blot. Three dilutions of antibody, 1:1000, 1:5000 and 1:10 000 was used to determine the optimal concentration. Membranes were cut prior to addition of the various anti-serum dilutions, approximate positions depicted by dotted lines, and visualised at the same time for 20 second exposure. The 1:1000 dilution was used for further experimentations.

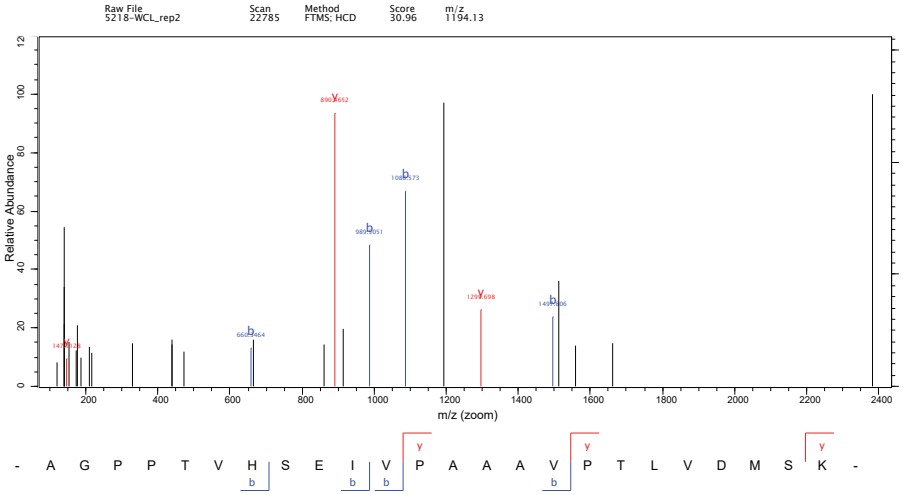
**Supplementary figure 7**



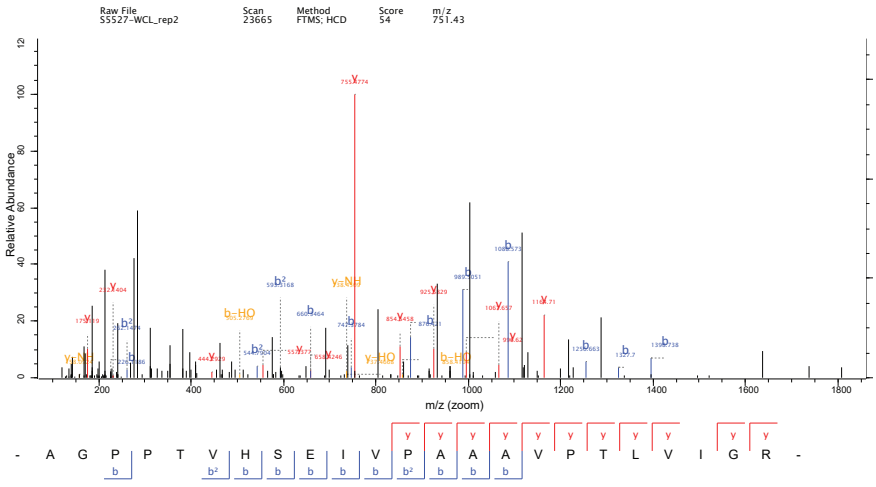

**Supplementary figure 7: Schematic illustration of a sequence of events executed computationally to filter structural variant files generated by our custom illumina pipeline (figure S2).** The code for this algorithm is written as an I/O operation in the bourne again shell language and implemented in the source code and can be executed from the main pipeline repository as an additional resource.

**Supplementary figure 8**

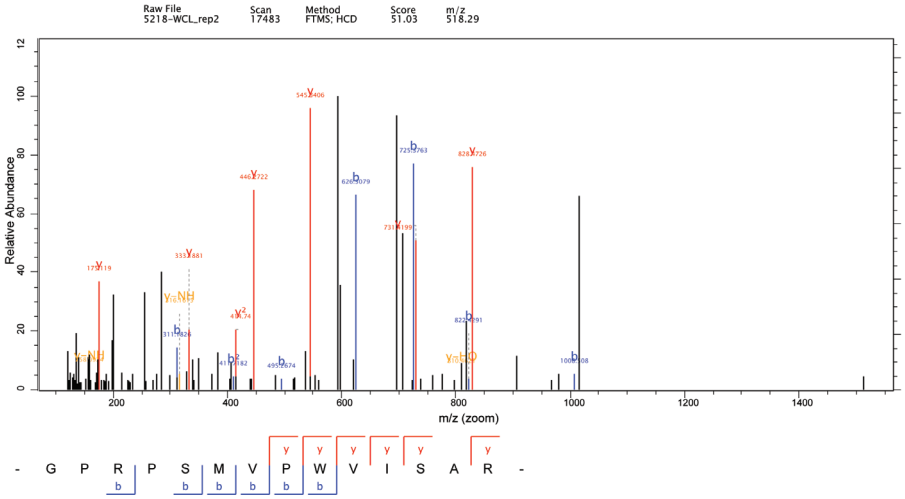
**A1**




**A2**



## Supplementary Figure 9




**Supplementary figure 9:** Tandem mass spectra assigned to a PKS15/1 fusion junction. Spectra was assigned using MaxQuant and was identified from whole cell lysates of *M. tuberculosis* S5218.

# 5

# PPE38-Secretion-dependent proteins of *M. tuberculosis* alter NF-kB signalling and inflammatory responses in macrophages

James Gallant[1,2]
Tiaan Heunis[1,3]
Caroline Beltran[1]
Karin Schildermans[4]
Sven Bruijns[5]
Inge Mertens[4]
Wilbert Bitter[2,6*]
Samantha L. Sampson[1*]

[1]DST/NRF Centre of Excellence in Biomedical Tuberculosis research, SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Science, Stellenbosch University, Tygerberg, 7505, Cape Town, South Africa
[2]Section Molecular Microbiology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam 1081 HZ, Amsterdam, The Netherlands
[3]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, United Kingdom
[4] VITO, Health Unit, Boeretang 200, 2400 Antwerp, Belgium
[5]Department of Molecular Cell Biology and Immunology, Cancer Center Amsterdam, Amsterdam Infection and Immunity Institute, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, Netherlands
[6]Medical Microbiology and Infection control, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam 1081 HZ, The Netherlands

Frontiers in Immunology

10.3389/fimmu.2021.702359

## ABSTRACT

It was previously shown that secretion of PE-PGRS and PPE-MPTR proteins is abolished in clinical *M. tuberculosis* isolates with a deletion in the *ppe38-71* operon, which is associated with increased virulence. Here we investigate the proteins dependent on PPE38 for their secretion and their role in the innate immune response using temporal proteomics and protein turnover analysis in a macrophage infection model. A decreased pro-inflammatory response was observed in macrophages infected with PPE38-deficient *M. tuberculosis* CDC1551 as compared to wild type bacteria. We could show that dampening of the pro-inflammatory response is associated with activation of a RelB/p50 pathway, while the canonical inflammatory pathway is active during infection with wild type *M. tuberculosis* CDC1551. These results indicate a molecular mechanism by which *M. tuberculosis* PE/PPE proteins controlled by PPE38 have an effect on modulating macrophage responses through NF-kB signalling.

**Keywords:** *Mycobacterium tuberculosis,* Proteomics, NF-KB signalling, PE/PPE, Macrophage

## INTRODUCTION

*Mycobacterium tuberculosis* is an important human pathogen that has adapted to survive and replicate within human macrophages (1). This lifestyle requires the presence of virulence factors that have evolved to enable intracellular growth. One strategy to identify virulence factors is by comparing the genomes of pathogenic and non-pathogenic mycobacteria. The most striking result of such an analysis is a large number of genes encoding PE and PPE proteins in *M. tuberculosis* compared to *M. smegmatis* (2). The PE and PPE proteins belong to two unique but related protein families and are known to be secreted to the cell surface or extracellular milieu through the type VII secretion system (3). The type VII secretion system is further represented by five loci in *M. tuberculosis,* named ESX-1 to ESX-5 (4). Of these five loci, ESX-2 and ESX-5 are the most recent, with ESX-5 being associated with the evolutionary split of the fast- and slow-growing mycobacteria (5). The PE/PPE proteins are characterised by the presence of a conserved N-terminal domain, including a Proline-Glutamic acid (PE) or a Proline-Proline-Glutamic acid (PPE) conserved motif (2). There are approximately 100 genes in *M. tuberculosis* coding for PE proteins and the major sub-family is classified as the PE-PGRS proteins, due to a large number of GC repeats in the genes encoding them (Polymorphic GC-Rich Sequence). This sub-family is characterised by multiple Gly-Gly-Ala/Gly-Gly-Asn repeats in the C-terminal region (6,7). The PE-PGRS proteins are unique to slow growing mycobacteria, localised to the cell surface and secreted through the ESX-5 system (4,8–11). Furthermore, recent studies demonstrated that the C-terminal region of PE-PGRS proteins can be cleaved from the PE domain and can be released to interact with the host (12). The largest subfamily of PPE proteins are the PPE-MPTR proteins (Major Polymorphic Tandem Repeat), which are also specific to slow-growing mycobacteria and secreted via ESX-5 (3). Taken together, the PE-PGRS and PE-MPTR proteins are interesting candidates for studying host-pathogen interactions considering their evolutionary history and extracellular localisation. However, elucidating the effects of PE-PGRS and PPE-MPTR proteins is hampered by the large number of proteins present in these groups. Recently, it has been shown that the *ppe38-71* operon, specifically the PPE38 protein, is involved in mediating secretion of both these important subfamilies and when deleted detectable secretion is abolished (13). More recent follow up studies conducted in *M. africanum* and *M. microti* demonstrated a lack of PE-PGRS secretion in the presence of an intact *ppe38-71* operon and ESX-5 secretion system (14,15). While it is clear that the *ppe38-ppe71* operon is involved in PE-PGRS secretion, this phenotype can be present in other mycobacteria, driven by an independent and currently unknown mechanism. By utilising a *ppe38*-71 mutant in *M. tuberculosis,* the collective role of the PE-PGRS and PPE-MPTR proteins that are dependent on both ESX-5 and PPE38 can be studied in the context of host-pathogen

interactions. Interestingly, a hypervirulent phenotype was observed with increased bacillary growth of an *M. tuberculosis ppe38-71* mutant in BALB/c mice (13). A similar phenotype was observed in zebrafish infected with a *ppe38* transposon mutant of *M. marinum* (13,16). It was further demonstrated that an *M. marinum ppe38* transposon mutant is able to modulate the innate immune response in murine macrophages and to alter antigen presentation (17). However, no differences were observed in TNF-α and CD40 levels in murine bone marrow-derived dendritic cells infected with an *M. tuberculosis Δppe38-71* mutant compared to wild type (18). The dispensability of *ppe38*, and by association the secretion of PE-PGRS proteins, is an unexpected observation as these proteins are hypothesised to be important for adaptation to an intracellular lifestyle.

To further explore the role of PE-PGRS and PPE-MPTR proteins in host-pathogen interactions, we characterised the temporal proteome profile of the THP-1 macrophage-like cell line in response to infection with *M. tuberculosis* CDC1551 and an isogenic *ppe38-71* mutant strain. We observed altered pro-inflammatory responses in macrophages infected with *M. tuberculosis Δppe38-71*. We further used stable isotope labelling of amino acids in cell culture (SILAC)-based proteomics to investigate protein turnover rates in response to infection and could show an increased turnover of proteins involved in pro-inflammatory responses in macrophages infected with *M. tuberculosis* CDC1551. Finally, our results suggest a role for PPE38-controlled PE/PPE proteins of *M. tuberculosis* in infection by modulating the inflammatory response through nuclear factor kappa B (NF-kB) signalling via the RelB pathway. By combining different approaches, we provide a deeper understanding of the molecular mechanisms exploited by *M. tuberculosis* to alter protective host responses during infection of macrophages.

## MATERIALS AND METHODS

### Bacterial strains

*Mycobacterium tuberculosis* CDC1551 *Δppe38-*(further referred to as *Δppe38-71*) and *M. tuberculosis* CDC1551 *Δppe38-71*::pMV_*ppe38-71* (further referred to as complemented) were generated from *M. tuberculosis* CDC1551 as parental strain with an integrated *ppe38-ppe71* serving as the complement under the constitutively expressed *hsp60* promotor. All three strains were obtained from a previous study (13). All bacterial strains were grown in modified Sauton's medium (0.4% L-asparagine, 0.4% glucose, 0.2% citric acid, 0.05% monopotassium phosphate, 0.05% magnesium sulphate, 0.005% ferric ammonium citrate and 0.001% zinc sulphate, pH 7.0), supplemented with 0.05% Tween-80 at 37°C without shaking in 75 $cm^2$ tissue culture flasks. *M. tuberculosis Δppe38-71* was

cultured in the presence of 50 µg/ml hygromycin (ThermoFisher, MA, USA) and *M. tuberculosis Δppe38-71*::pM*V_ppe38-71* was cultured in the presence of 25 µg/ml kanamycin (Sigma-Aldrich, MO, USA) and 50 µg/ml hygromycin as indicated. Antibiotics were only used during pre-culture and were omitted during sub-culturing. *In vitro* growth was monitored by culturing *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain as stated above and measuring optical density at 600 nm over 20 days.

The mean was derived from four (n = 4) biologically independent experiments and error bars represent standard error of the mean (SEM). Statistical differences between strains at each time point was determined by two-way ANOVA followed by a Tukey HSD post-hoc test. The q-value used to infer statistically significant difference was set at 0.05, thus only a q-value below this number was considered significant. The statistical tests were performed using the R statistical programming language version 3.8.1 using the standard library. See Fig. 1B and the corresponding legend for detailed statistical descriptions related to each figure.



## Cell lines

The THP-1 human monocytic cell line (ATCC®TIB-202©) available from the American Type Culture Collection (ATCC, VA, USA) was stored as frozen seed lots at -80°C suspended in cell freezing media (Merck, NJ, USA) until use. The THP-1 cells were cultured in RPMI (ThermoFisher, MA, USA) media supplemented with 10% fetal bovine serum (FBS) (ThermoFisher, MA, USA), further referred to as R10 media, at 37°C and in a 5% $CO_2$ atmosphere for a maximum of three passages in either 25 or 75 $cm^2$ tissue culture flasks. For pulse-chase SILAC (pSILAC) experiments, the THP-1 monocytes were grown as described above, however, the 10% FBS was substituted for 10% dialysed FBS ($R10_{LIGHT}$). THP-1 monocytes were differentiated into macrophage-like cells using 50 ng/ml phorbol myristate acetate (PMA) (Sigma-Aldrich, MO, USA) for three days. For experiments in *ex vivo* bacterial growth, 48-well tissue culture grade plates were used with 1 x $10^5$ THP-1 seeding density. Proteomics experiments involving THP-1 cells used six well plates with a seeding density of 1 x $10^6$ cells. The media was changed after three days, the cells washed twice with phosphate-buffered saline (PBS) pH 7.4 (ThermoFisher, MA, USA) and the media replaced with either R10 or $R10_{LIGHT}$. The THP-1 macrophage-like cells were subsequently rested for 24 hours at 37°C in a 5% $CO_2$ atmosphere before performing further experiments.

## Macrophage infection

*M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain were grown, separately, as described above for 4 days or when an $OD_{600}$ of 1.0 was reached. The cells were harvested by centrifugation (4 000 rpm, 10 min) and washed three times with

PBS pH 7.4 to remove Tween-80. Mycobacterial cells were subsequently resuspended in 5 ml PBS pH 7.4 and sonicated for 10 minutes in a water bath sonicator, followed by filtering through a 40 µm cell strainer to remove bacterial clumps. Optical density (OD) was measured, and suspensions were diluted to an $OD_{600}$ of 0.1 (corresponding to approximately 1 x $10^7$ cells/ml) in R10 medium.

The THP-1 macrophage-like cells were infected with the different *M. tuberculosis* strains at a multiplicity of infection (MOI) of 3:1 and incubated for 3 hours at 37°C in a 5% $CO_2$ atmosphere to allow for uptake of mycobacteria. For label-free proteomics, the infected macrophages were washed seven times with PBS pH 7.4 to remove extracellular bacteria and proteins were harvested at 4 hours and 18 hours post-infection, as described below. For colony forming unit (CFU) determination, the infected macrophages were treated with 100 U/ml pen-strep (penicillin/streptomycin) at 37°C for 1 hour, followed by three washes with PBS pH 7.4. Macrophages were lysed using deionised water at the indicated time points. The lysates were serially diluted to a maximum dilution of 1 x $10^{-6}$ and plated on 7H11 agar, followed by incubation at 37°C for ~3 weeks or until colonies formed. To account for the presence of extracellular bacteria, THP-1 macrophage-like cells were washed seven times with PBS pH 7.4 to remove mycobacteria that were not internalised after three hours of infection. Bacteria remaining after the consecutive washes were determined by plating the final wash on 7H11 agar and incubated at 37°C for ~3 weeks or until colonies formed.

The bacterial viability and growth within macrophages were determined from three independent experiments described in Fig. 1C, Fig. 1D, Fig. 1E and the corresponding legend. Data was gathered by infecting THP-1 macrophage-like cells with mycobacterial strains and enumerated by colony forming unit (CFU) counts. The results were derived from three biologically independent experiments (n = 3) and error bars represent SEM. Hypothesis testing was performed in the R statistical programming language version 3.8.1 using two-way ANOVA with a Tukey HSD post hoc test for data represented in supplementary Fig. 1C and 1E while a one-way ANOVA with a Tukey HSD post hoc test was used for Fig. 1D. Statistical significance was inferred using the resulting q-value and a cut-off was set at 0.05.

For pSILAC experiments, THP-1 macrophage-like cells were washed three times with PBS pH 7.4 and the media replaced with SILAC RPMI devoid of lysine and arginine (ThermoFisher, MA, USA). The cells were subsequently incubated for 3 hours at 37°C in a 5% $CO_2$ atmosphere prior to infection to facilitate amino acid uptake upon exposure to mycobacteria. After the three-hour incubation, the SILAC RPMI was replaced with fresh SILAC RPMI supplemented with 10% dialysed FBS (ThermoFisher, MA, USA), 0.4

mg/ml $^{15}N_4{}^{13}C_6$-arginine and 0.08 mg/ml $^{15}N_2{}^{13}C_6$-lysine ($R10_{HEAVY}$) and placed on ice until infection.

*M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain were cultured and prepared for infection as previously described, with minor modifications. Briefly, the mycobacterial cells were harvested, sonicated and filtered as described above, however, the cells were diluted to an $OD_{600}$ of 0.1 in $R10_{HEAVY}$ medium. Macrophage infections were further carried out as described above. Proteins were harvested at 4 hours, 8 hours, 12 hours, 18 hours and 28 hours post-infection for protein turnover analysis, as described below.

## MTT assay

The 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) reagent was prepared in $R10_{Light}$ to a final concentration of 5 mg/ml. Supernatants of infected as well as uninfected macrophages were removed following infection at 18 hours and replaced with $R10_{Light}$-MTT media. The macrophages were subsequently incubated at 37°C, 5% $CO_2$ atmosphere for two hours. The media containing MTT was removed after this incubation step and replaced with DMSO and incubated at 37°C, 5% $CO_2$ atmosphere for an additional 15 minutes. Macrophage viability was calculated by colorimetric shift measured at 540 nm using a plate reader.

The viability assay is described in Fig. 1G, the corresponding legend and details of the statistical test can be found in Data S1: Table S1. Results were gathered by infecting THP-1 macrophage-like cells with each mycobacterial strain as well as an uninfected control and measuring optical density at 18 hours post infection. The data in this experiment was derived from three (n = 3) independent experiments and significant differences were detected by one-way ANOVA following a Tukey HSD post hoc test. This test was performed in the R statistical programming language version 3.8.1 using the standard library and a q-value less than 0.05 was considered significant.

## Protein extraction and proteomic sample preparation

At each time point, infected THP-1 macrophage-like cells were washed four times with PBS pH 7.4 on ice. Modified RIPA buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1mM EDTA, 1% IGEPAL CA-630, 0.1% sodium deoxycholate, Benzonase and Roche Complete EDTA-free protease inhibitor) was added to each well and placed on ice for 10 minutes to lyse cells. The lysates were subsequently centrifuged (14 000 rpm, 20 min) to remove cellular debris. Proteins were precipitated by the addition of four volumes ice-cold acetone and quantified by modified Bradford assay (19).

All bacterial strains were grown in modified Sauton's medium (0.4% L-asparagine, 0.4% glucose, 0.2% citric acid, 0.05% monopotassium phosphate, 0.05% magnesium sulphate, 0.005% ferric ammonium citrate and 0.001% zinc sulphate, pH 7.0), supplemented with 0.05% Tween-80 at 37°C without shaking. For Tween-free supernatants, the bacteria was grown in tween until an OD of ~1 was reached and Tween-80 was removed by sequential washing through centrifugation (4 000 rpm, 20 minutes) for a total of 4 washes and sub-cultured in Sautons media without Tween-80. The bacteria was cultivated for an additional 5 days at 37°C before harvesting. Bacterial whole-cell lysates were produced by separating the cells from the supernatant by centrifugation (4 000 rpm, 4°C, 10 minutes). The supernatant was filter sterilised using 0.22 µm low bind syringe filters (Merck Milipore, NJ, USA) and used for the Tween-80 supernatant fraction. The pellet was resuspended in 1 ml ice cold protein extraction buffer (10 mM Tris-HCl pH 7, 0.1% Tween-80, Complete Protease inhibitor cocktail table) and centrifuged once more (14 000 rpm, 4°C, 2 minutes). The pellet was suspended in 300 µl cold protein extraction buffer supplemented with RNase-free DNaseI and transferred to cryogenic vials containing an equal volume 0.1 mm acid washed glass beads. The suspension was mechanically lysed by bead-beating for 30 second intervals with 30 seconds on ice for a total of 8 cycles. The resulting lysate was clarified by centrifugation (14 000 rpm, 4°C, 15 minutes) and filter sterilised through 0.22 µm pore acrodisks and syringes. Tween-free supernatants were harvested by separating the cells from the supernatant using 40 µm cell strainers (Corning, NY, USA) followed by centrifugation (4 000 rpm, 10 minutes) and filter sterilised using 0.22 µm Steriflip® filters (Sigma-Aldrich, MI, USA). The supernatants were concentrated using Amicon Ultra 3 KDa spin columns and precipitated with ice cold acetone overnight. The precipitated proteins were resuspended in 8M urea, and all proteins were quantified using a modified Bradford assay and used for LC-MS/MS (19).

Equal amounts of protein (20 µg) were used for sample preparation for LC-MS/MS following a tube gel protocol (20). Briefly, proteins suspended in 8M urea were diluted in 1M tris-HCl pH 6.8 and cast in an acrylamide gel containing 10% SDS within an Eppendorf tube. The samples were allowed to set overnight at room temperature. Protein-containing gels were removed from the Eppendorf tubes and cut into ~1 mm x 1 mm gel pieces. Gel pieces were washed three times with 50 mM ammonium bicarbonate (Sigma-Aldrich), followed by reduction with 5 mM tris(2-carboxyethyl)phosphine (Sigma-Aldrich, MO, USA) for 1 hour at 45°C. After reduction, the proteins were alkylated with 55 mM iodoacetamide (Sigma-Aldrich, MO, USA) for 1 hour at room temperature. The gel pieces were washed twice with 100% acetonitrile after alkylation. Sequencing grade modified trypsin (Promega, WI, USA) was added to the gel pieces at a 1:50 trypsin to protein ratio and incubated at 4°C for 1 hour, followed by an 18h proteolytic digestion

at 37°C in a humidified chamber. Peptides were eluted by sequential addition of 50%, 70% and 100% acetonitrile until gel pieces turned opaque. The eluates were then dried in a vacuum desiccator (SpeedVac). Dried peptides were suspended in 5% acetonitrile (Sigma-Aldrich, MO, USA) containing 0.1% formic acid and desalted using C18 desalting columns (ThermoFisher, MA, USA) as recommended by the manufacturer.

## LC-MS/MS

Dried peptides were dissolved in 30 µl of solvent A (2% acetonitrile containing 0.1% formic acid in HPLC-grade water) and 500 ng peptides was analysed. All chromatography was performed on a nanoAcquity UPLC system (Waters, MA, USA) using a 200 cm uPAK™ column (Pharmafluidics, Gent, Belgium) coupled to a Thermo Q-Exactive Plus Orbitrap mass spectrometer (ThermoFisher, MA, USA) equipped with a Flex nano-electrospray source. The spray voltage was set to 1.9 kV (Thermo Fisher, MA, USA) and the capillary temperature was 250°C. Peptide separation was performed using a linear gradient of solvent B (98% acetonitrile, 0.1% formic acid and 2% HPLC grade water), starting with 3% solvent B and increased to 40% solvent B over 80 minutes. Solvent B was increased to 100% in 5 minutes and subsequently decreased to 3% solvent B in 5 minutes. Solvent B was kept at 3% B for an additional 35 minutes at a flow rate of 750 nL/min, where a column temperature of 50°C was maintained with a heater. Mass spectrometry was performed in data-dependent acquisition mode using a full MS1 scan (350-1 850 m/z, resolution at 70 000, max injection time was 100 ms and ACG target was 3e6), and selecting precursor ions with a $2^+$ or greater charge state for MS/MS analysis. This was followed by HCD fragmentation with normalised collision energy set at 28% and MS/MS acquisition (200–2 000 m/z, resolution 17 500, max injection time of 80 ms, AGC target was 1e5) of the top 20 most intense precursors from each full scan. Dynamic exclusion of ions was implemented using a 20s exclusion duration and only ions with an unassigned charge state were disregarded.

## Mass spectrometry data analysis

All tandem mass spectra were analysed using MaxQuant version 1.6.10 (21) and searched against the human proteome (UP000005640, containing 74 349 entries) database downloaded on 14/10/2017 from Uniprot and the *M. tuberculosis* CDC1551/Oshkosh proteome (UP000001020, containing 4 204 entries) downloaded on 17/4/2017. Peak list generation of label-free tandem mass spectra was performed within MaxQuant using default parameters and the built-in Andromeda search engine (22). Enzyme specificity was set to consider fully tryptic peptides with two missed cleavages were allowed. Oxidation of methionine and N-terminal acetylation were allowed as variable modifications. Carbamidomethylation of cysteine was allowed as a fixed modification. A protein and peptide false discovery rate of less than 1% was employed in MaxQuant

with match between runs enabled. Proteins that contained similar peptides that could not be differentiated on the basis of MS/MS analysis alone were grouped to satisfy the principles of parsimony. Data handling, statistical tests and figure generation was performed using ProVision, an online data analysis platform that uses the LIMMA package in R for statistical tests (23). Briefly, Reverse database hits, contaminants and proteins only identified by site modifications were removed. Precursor intensity values for each protein was obtained from MaxQuant using the MaxLFQ algorithm available internally (24). The file was further filtered for each protein group to contain at least two unique peptides. The assigned LFQ intensity values were subsequently log2 transformed to gain a normal distribution and further filtered for two values in at least one group. This resulted in the high confidence expression dataset, and missing values were imputed from a truncated normal distribution of transformed LFQ intensities. Quantile of Quantile plots were used within the ProVision application to check for normality prior to statistical testing. Multiple hypothesis testing was corrected using the Benjamini-Hochberg FDR set at 0.05, and a two-fold cut-off was implemented. The statistical analysis and visualisations of label-free mass spectrometry data can be found in supplementary Fig. 1 and Fig. 2 with an extended analysis in supplementary Fig. S2. The mass spectrometry data pertaining to this experiment, which was used to generate these figures are available in Data S1, table S2. Differential expression of *in vitro* grown mycobacterial strains was as described in Fig. 5E, 5F, supplementary Fig. 5A and supplementary Fig. 5B as well as Data S1: table S5, table S6 and table S7. The data used for hypothesis testing was derived from three (n = 3) biologically independent experiments for each condition. Significant differences were accepted when the q-value was below 0.05 and a log2 fold change of 1.

## Protein turnover

The search parameters for SILAC were largely similar to that of the label-free searches using MaxQuant version 1.6.10 and the same human proteome (UP000005640) as a reference database. For the pSILAC searches a multiplicity of 2 was chosen adding heavy arginine with a mass shift of 10 Da and heavy lysine with a mass shift of 8 Da. Furthermore, the re-quantify function was disabled and match between runs was enabled. Oxidation of methionine and N-terminal acetylation was chosen as variable modification and carbidomethylation of cysteine was set as the fixed modification and all FDR cut offs for peptide identification was set at 0.01. Under conditions with increased arginine cell lines can convert the excess arginine to proline which results in non-specific dilution in the stable isotope signal and can confound intensity values. We therefore tested this conversion in our differentiated THP-1 macrophage-like cell line with LPS stimulating with heavy lysine and arginine for 18 hours. Searching in the same manner as above with proline (6 Da shift) as a variable modification only iden-

tified 5 proteins containing this amino acid. Thus, we continued with the 0.4 mg/ml $^{15}N_4{}^{13}C_6$-arginine for the rest of our experiments. The resulting protein groups file was filtered in the same manner as detailed above for the label-free proteomics. Briefly, all contaminants; reverse database hits as well as proteins only identified by site modifications was removed from the SILAC dataset and we additionally filtered for a minimum of two unique peptides. The post-filtered raw ratios of each strain and each time point was used for further analysis.

As the heavy isotopes increase, the light isotopes decay providing a function to extrapolate half-life for the population of proteins based on linear regression. Protein turnover was calculated from an average from the raw ratios using a similar approach as previously described (25). As the natural occurrence of $^{15}N_2{}^{13}C_6$-lysine is exceedingly rare we assumed a zero time point where the heavy fraction was set to zero for further handling in a manner as previously described (26). The raw ratios of each condition (i.e. THP-1 infected with CDC1551, *Δppe38-71* etc.) were filtered to contain at least three ratios in the time course and with a coefficient of determination above 0.85. The half-lives ($T_{1/2}$) of each protein were determined using a first order reaction equation (eq 1). This was subsequently derived for calculating the half life (eq 2), where half-life is equal to the natural log of two divided by the rate ($K_{dp}$)

$$N_t = N_0 e^{-K_{dp} t} \tag{1}$$

$$T_{\frac{1}{2}} = \frac{ln(2)}{K_{dp}} \tag{2}$$

The reaction rate ($K_{dp}$) was calculated using equation 3 without accounting for dilution by replication as THP-1 macrophage-like cells are terminally differentiated. Here $r$ is representing the raw ratio at each time point and $t_i$ is each specific time point. For optimal handling the equation can be expressed using all variables for a direct calculation of half-life (eq 4).

$$K_{dp} = \frac{\sum ln(r+1)t_i}{\sum {t_i}^2} \tag{3}$$

$$T_{\frac{1}{2}} = ln(2) \div \left(\frac{\sum ln(r+1)t_i}{\sum {t_i}^2}\right) \tag{4}$$

Calculating the $K_{dp}$ values were implemented using the Excel macro LinEstGap which implements equation 3 (see data availability for link to the macro) and the half-lives were calculated using equation 4. The average half-life was calculated after filtering for each replicate and was further processed in the R statistical programming language. Here the average half-life was calculated per condition which produced a log normal

distribution and missing values. This was followed by filtering the dataset where a protein should contain at least one half-life representing at least one of the four conditions.

For further analysis we Z-scored the half-lives to obtain a dataset used to generate heatmaps where the row clustering was to gain a distance from the centroid metrics. These clusters were enriched for gene KEGG pathways using the WebgestaltR package available on the CRAN repository as shown in Fig. 3D (27). Hypothesis testing was used to determine differentially regulated half-lives from two biologically independent experiments (n = 2) across five time points namely 4 hours, 8 hours, 12 hours, 18 hours and 28 hours post infection. The difference in protein turnover between CDC1551 and *Δppe38-71* infected THP-1 macrophage-like cells was determined using hypothesis testing. This allowed for stringent filtering, i.e. contains ratios in both replicates in both conditions, without losing valuable data. The strict filtered dataset for each comparison was log2 transformed to obtain a normal distribution and the Limma package was used for hypothesis testing. The resulting p-values were corrected for using Benjamini-Hotchberg FDR set at 0.05 and fold change was disregarded for these specific analyses. The results of these tests were visualised in Fig. 4A and Fig. 4B, the raw data used to generate these figures are available in Data S1: table S4. The differentially regulated proteins were further used to enrich for KEGG pathways using the WebGestaltR package as shown in Fig. 4D.

## Western blots

Macrophage whole-cell lysates and bacterial supernatants were either probed with a WesternBreeze anti-rabbit chemiluminescent kit (Thermo Fisher, MA, USA) or manually if primary antibodies required an anti-mouse secondary antibody. Briefly, whole-cell lysates were quantified using a modified Bradford assay and 50 µg total protein content was separated by SDS-PAGE. Proteins were subsequently transferred to nitrocellulose membranes for 1 h. The membranes were stained with Ponceau S (0.1%, w/v in 5% acetic acid) to inspect sample loading. Membranes were probed with anti-GAPDH (CST, D16H11), anti-IL-1B (CST; D3U3E), anti-NF-kB1 (CST; D4P4D) anti-NF-kB2 (CST,18D10), anti-ISG15 (CST; 22D2), anti-DDX58 (CST; D14G6), anti-PPE38 (custom, Innovagen, Sweden) or anti-PGRS (3) overnight at 4°C, as indicated in the text. This was followed by incubation with an alkaline phosphatase-conjugated secondary anti-rabbit antibody or a horseradish peroxidase-conjugated goat anti-mouse secondary antibody for 1 hour at room temperature. Blots were visualised using the chemiluminescent substrate provided by the WesternBreeze kit or with ECL detection reagent (Bio-Rad) and imaged on a ChemiDoc Imaging System (Bio-Rad).

## Macrophage stimulations

*M. tuberculosis* CDC1551, *Δppe38-71* and complemented strains were grown in Sauton's media containing 0.05% Tween-80 for 5 days or until an $OD_{600}$ of ~1 was reached. *M. tuberculosis* culture supernatants were harvested by centrifugation (4 000 rpm, 10 min) and concentrated using 4 kDa molecular weight cut off Amicon Ultra spin columns (Merck Millipore, MA, USA). Proteins in the cell-free supernatants were precipitated using ice-cold acetone and kept overnight at -20°C. The precipitated proteins were harvested by centrifugation (14 000 rpm, 30 min) and resuspended in 8M urea in 50 mM triethylammonium bicarbonate (urea buffer). Protein concentration was determined using a modified Bradford assay, as described previously (19).

THP-1 monocytes were seeded into 6-well plates at 1 x $10^6$ cells/well and differentiated into macrophages, as described above. THP-1 macrophage-like cells were stimulated for 18h with lipopolysaccharide from *E. coli* (Sigma-Aldrich, MO, USA) at 100 ng/ml or proteins from cell-free supernatants of *M. tuberculosis* at 20 µg per well. Unstimulated cells served as baseline controls for IL-1B expression. Whole-cell lysates of the macrophages were harvested as described above and stored at -80°C until further use.



Expression was quantified by densitometry of western blots, where a represented blot is depicted in supplementary Fig. 5C. Relative quantification was determined from three independent biological experiments (n = 3) and depicted in supplementary Fig. 5D. Significant differences were determined by one-way ANOVA followed by Tukey HSD and details are depicted in supplementary Fig. 5E. Significant differences were accepted when the q-value was below 0.05.

## Immunofluorescent staining and microscopy

THP-1 monocytes were cultivated, differentiated and infected as described above. After 18 hours of infection, the medium was removed, the cells were fixed in 4% paraformaldehyde (PFA) for 30 minutes at room temperature and washed three times in PBS pH 7.4 for 10 minutes. Cells were permeabilised using 0.1% Triton X-100 in PBS pH 7.4 for 10 minutes at room temperature and washed three times in PBS pH 7.4 for 10 minutes. Cells were blocked with 1% bovine serum albumin (BSA) in PBS pH 7.4 containing 0.1% Tween-80 and 0.1M glycine for 1 hour at room temperature, after which cells were incubated with corresponding primary antibodies diluted in 1% BSA in PBS pH 7.4 overnight at 4°C in a humidified chamber. Antibody dilutions were used as follows: RelA antibody (NF-kb p65 (D14E12) XP® Rabbit mAb, Cell Signalling Technology, MA, USA) was used at a 1:400 dilution; RelB (Recombinant Anti-Rel B antibody [EPR7076] - C-terminal (ab180127, Abcam, Cambridge, UK) was used at a 1:200 dilution; NF-kb1 p105/p50 (D4P4D, Rabbit mAb, Cell Signalling Technology, MA, USA) was used

at a 1:200 dilution; NF-kb2 p100/p52 (D7AK9, Rabbit mAb, Cell Signalling Technology, MA, USA) was used at a 1:400 dilution. Cells were washed three times in PBS pH 7.4 for 10 minutes and incubated with anti-rabbit Alexa Fluor 488 secondary antibody (Thermo Fisher, MA, USA) used at 1:500 dilution for 1 hour at room temperature in the dark. Cells were washed three times in PBS pH 7.4 for 10 minutes and subsequently stained with Phalloidin–Tetramethylrhodamine B isothiocyanate (Sigma Aldrich, MI, USA) to visualize F-actin. Nuclei were counterstained with Hoechst in PBS pH 7.4 for 10 minutes at room temperature in the dark. Cells were washed a further three times in PBS pH 7.4 for 10 minutes before mounting upside down onto glass microscope slides using Dako fluorescent mounting medium and air drying overnight in the dark at room temperature. Slides were stored at 4°C in the dark until imaging. Unstained, single stained and secondary antibody only controls were prepared for each experiment to assess background autofluorescence and signal specificity in each channel. Images were obtained using a Carl Zeiss LSM 780 confocal microscope (Plan-Apochromat x63/1.40 oil DIC M27 objective lens). Images were acquired using the ZEN software (Carl Zeiss, Oberkochen, Germany). Acquisition settings for imaging were identically set for all treatment groups within each experiment. Analysis and quantification of images were done with the ZEN black software version 13.0.0.518 (Carl Zeiss, Oberkochen, Germany) and any manipulations done, such as min/max, were extended to all channels. No single channel enhancements were used during quantification and all changes were applied equally across the entire image. In addition, twenty cells were chosen from at least 10 random fields per replicate to with a minimum of three independent experiments (n=3) to generate the results as depicted in Fig. 7J. For display images, channels were enhanced as appropriate to accurately demonstrate nuclear translocation of proteins.

## ELISA

Supernatants harvested from infected THP-1 macrophage-like cells at 18 hours and 48 hours post-infection were sterilised using a 0.22 µm syringe filter. Sterilised supernatants were subsequently assayed for interleukin 12 p70 (IL-12p70) and interleukin 3 (IL-13) levels using ELISA kits (Biosource, Invitrogen), as indicated by the manufacturer.

The data was derived from three independent biological experiments (n = 3) and is depicted in Fig. 8A, Fig. 8B and details of the statistical test can be found in Fig. 8 legend and Data SI Table S8. Significant differences were determined by one-way ANOVA followed by Tukey HSD and accepted when the q-value was below 0.05.

# RESULTS

## PPE38 does not influence *M. tuberculosis* uptake or replication within macrophages

The cell surface of *M. tuberculosis* is covered with pathogen-associated molecular patterns (PAMPs), which will be recognised by the host during infection to initiate immune responses. The PE-PGRS and PPE-MPTR proteins are groups of PAMPs localised on the cell surface of *M. tuberculosis* that have been implicated in host-pathogen interactions and virulence (28–30). Previous work has linked a *Δppe38-71* deletion mutant to increased virulence in *M. tuberculosis* by controlling a subset of PE-PGRS and PPE-MPTR secretion (13).

In the present study, we exploited a *Δppe38-71* deletion mutant to test whether a lack of PPE38 and its effectors will result in altered macrophage responses during infection. Initially, we verified that the strains used in this study have the same phenotype of diminished PE-PGRS protein secretion previously observed (13), as well as verifying the presence and absence of PPE38 by western blot (Fig. 1A). Next, we measured the growth of *M. tuberculosis* CDC1551, the *Δppe38-71* mutant and the complemented strain to determine if any growth differences exist. No significant differences were found between the strains over a 20-day growth period (Fig. 1B). To reduce variability, 4-day old mycobacterial cultures were used for all infections. The growth of each strain was monitored within THP-1 macrophage-like cells for 28 hours by enumerating cell counts at various time points (4 hours, 18 hours and 28 hours) post-infection. No significant differences were found in mycobacterial proliferation or macrophage cell death associated with any of the strains over the 28-hour period (Fig. 1C). For the mass spectrometry experiments, we omitted antibiotic treatments for the killing of extracellular bacteria, which could result in confounding factors in this experimental design. Instead, extracellular bacteria were removed by multiple washes. The number of remaining bacteria was enumerated by CFU and a two log decrease of extracellular bacteria was observed (Fig. 1D). Mycobacterial uptake was measured by plating intracellular bacteria at 4 hours, 18 hours and 28 hours to rule out uptake of the residual extracellular mycobacteria during the incubation period. The intracellular bacterial load remained steady across the time points and no significant differences were detected, thereby indicating little to no ongoing uptake within the time frame of this experiment (Fig. 1E). The percentage uptake from the original titre at 4 hours after infection was 10-15% for all strains, with no significant difference between strains (Fig. 1F). Lastly, an MTT assay was used to determine macrophage viability at 18 hours post-infection. A decrease in THP-1 viability between infected and uninfected states was observed (Fig. 1G). However, no statistically significant differences in cell death were observed

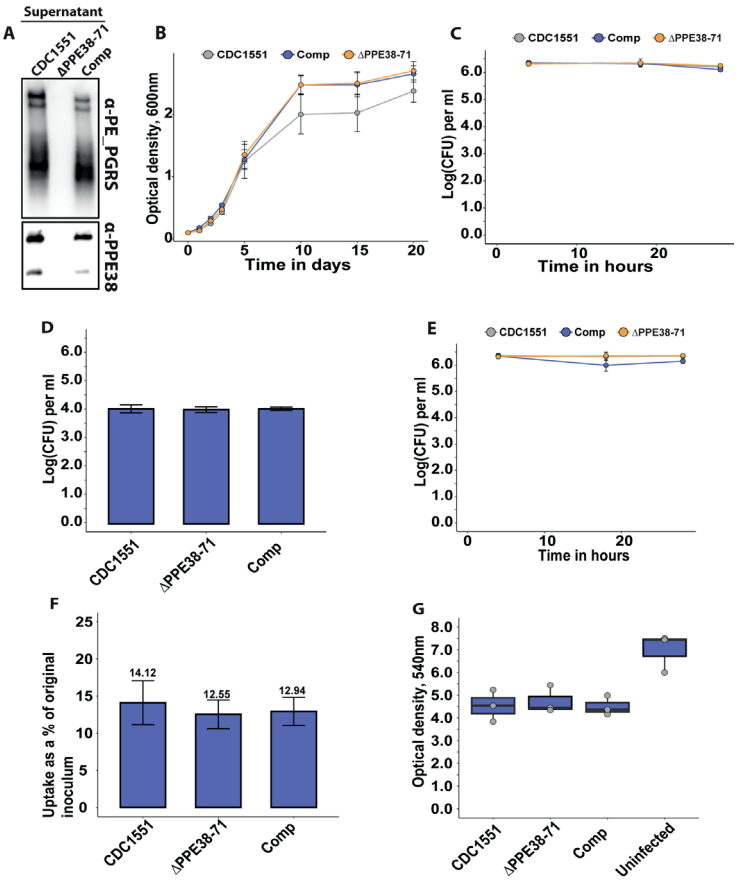



**Figure 1: The absence of PPE38 controlled PE-PGRS proteins has no effect on bacterial survival or macrophage death in a 28 hour period.**

**A)** PE-PGRS and PPE38 immunoblot of *in vitro* culture supernatants of *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain harvested at $OD_{600}$ of ~1. **B)** *In vitro* growth curves of *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain by optical density measurements for 20 days. Data is representative of four independent experiments and error bars represent SEM. Two-way ANOVA with Tukey HSD post-hoc testing was used to determine a significant difference with a q-value cut-off at 0.05 **C)** Intracellular growth curve of *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain in THP-1 macrophage-like cells. Growth was monitored over 28 hours and infected macrophages were harvested at 4 hours, 18 hours and 28 hours after exposure to mycobacteria. Mycobacterial load was measured by CFU enumeration. Data is representative of three independent experiments and error bars represent SEM. **D)** Extracellular bacteria at 3 hours post-infection were determined by CFU enumeration after 7 consecutive washes. No significant difference was detected by one-way ANOVA between the different strains. Data is representative of three independent experiments and error bars are indicative of SEM. **E)** Differential uptake of mycobacteria by THP-1 macrophage-like cells during infection was measured by washing excess bacilli at 3 hours post-infection. Pen/Strep treatment for 1 hour was used directly before each harvesting time point, i.e. 4 hours; 18 hours and 28 hours; to remove any extracellular bacteria still present. The intracellular mycobacterial load was determined by CFU enumeration and data is representative of three independent experiments, error bars represent SEM. **F)** Percent bacterial uptake calculated from the original inoculum at an MOI of 3:1. Data is representative of three independent experiments where the percentage was calculated from the mean CFU of each strain obtained from 4 hours post-infection divided by the mean CFU of the original inoculum. **G)** THP-1 macrophage viability was measured by MTT assay at 18 hours post-infection. Data is representative of three independent experiments and error bars represent SEM. One-way ANOVA was used to determine significant differences with a p-value cut off of 0.05.

in macrophages infected with the different strains (Data S1, Table S1). These experiments demonstrate similar growth of *M. tuberculosis Δppe38-71* and wild type *in vitro*. The *Δppe38-71* mutation does not alter macrophage viability compared to wild type, nor is intracellular survival of *M. tuberculosis Δppe38-71* affected compared to wild type in our experimental conditions. These experiments indicated that abrogation of PE-PGRS/PPE-MPTR secretion to the extracellular milieu does not cause the macrophage to clear the mycobacterial infection or succumb to it. However, given the immunogenic nature of the PE-PGRS and PPE-MPTR proteins some effect is to be expected upon loss of their secretion (28,29). We therefore investigated the proteome response of THP-1 macrophage-like cells to infection with *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain.

## Label-free proteomics reveals time-dependent differences in pro-inflammatory responses in macrophages when exposed to *M. tuberculosis* strains lacking PPE38-and PPE38 controlled proteins.



Whole-cell lysates were harvested from infected macrophages at 4 hours and 18 hours post-infection and analysed by label-free mass spectrometry. A total of 2 052 confident protein groups (FDR < 1%) were identified after filtering for two unique peptides and a minimum of two values per replicate in at least one of the groups. The label-free quantification (LFQ) algorithm in MaxQuant was used for relative protein quantification, and no differences in the mean distribution of intensity values were observed (Fig. S1A). Principal component analysis (PCA) of LFQ intensities was used to determine clustering within and between replicates of each test group at each time point. No distinct clustering was observed at the 4h time point (Fig. S1B), indicating that early innate immune responses are similar between macrophages infected with the different strains. However, a clear separation of the groups could be observed at the 18-hour time point in the first component, indicating that the majority of the variation is due to strain-specific features (Fig. S1C). The same analysis of the temporal macrophage responses within the groups at different time points also revealed separation in the first component (Fig. S1D, S1E, S1F).

Next, we determined the differentially regulated proteins in the infected THP-1 macrophage-like cells. There was no significant difference in protein expression between *M. tuberculosis* CDC1551 and *Δppe38-71* infected macrophages at 4 hours post-infection (Fig. 2A). At 18 hours post-infection, a total of 39 proteins were downregulated in macrophages infected with *M. tuberculosis Δppe38-71* compared to *M. tuberculosis* CDC1551, while 11 proteins were upregulated (Fig. 2B). Furthermore, macrophages infected with *M. tuberculosis* CDC1551 had an increased temporal cytokine expression profile, where the majority of the 39 upregulated proteins were involved in the expression of proteins

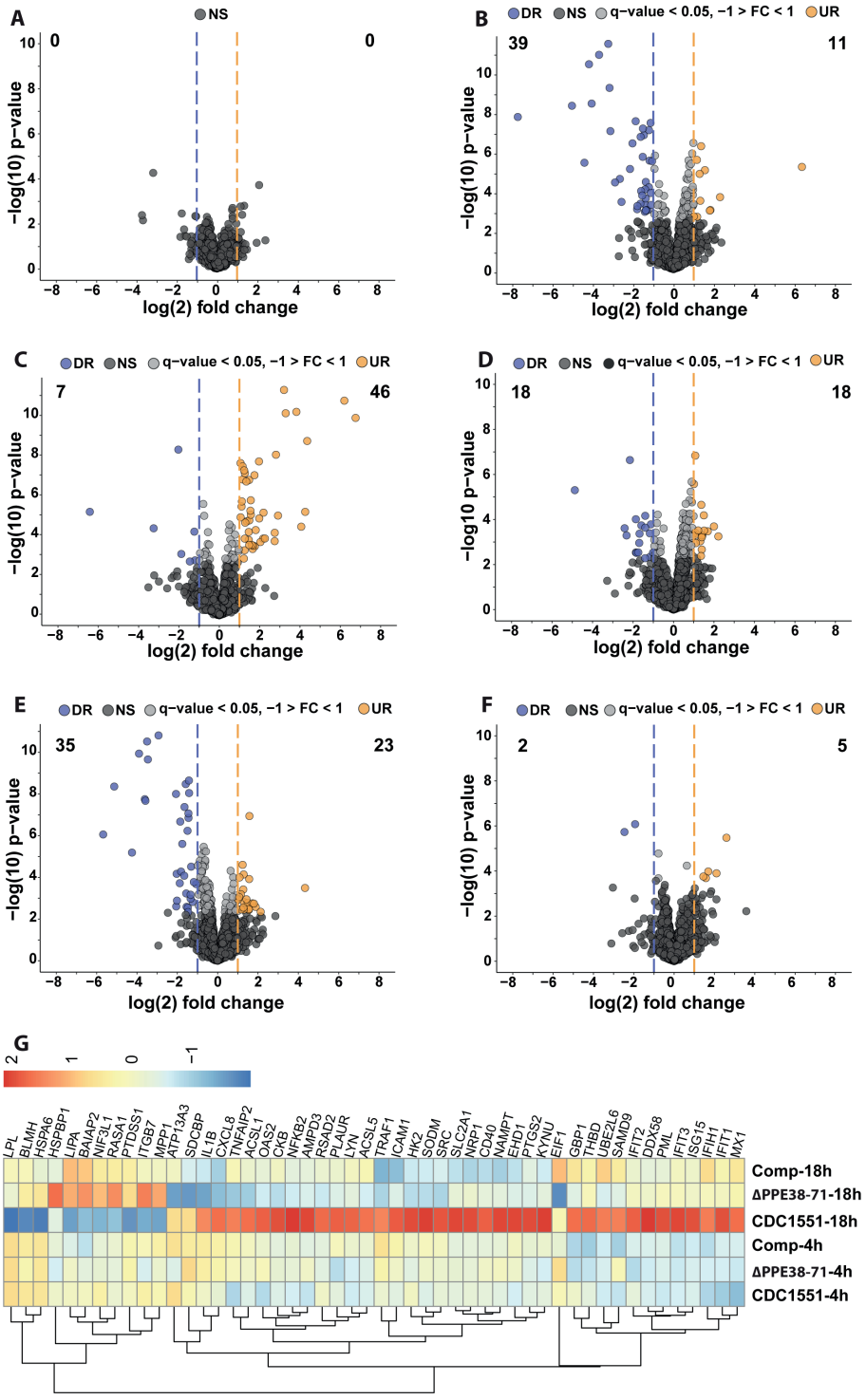


**Figure 2. Label-free proteomic analysis reveals altered inflammatory responses in *M. tuberculosis*-infected macrophages.**
Volcano plots representing differential protein abundance in macrophages infected with *M. tuberculosis* CDC1551 compared to infection with *Δppe38-71* at **A)** 4 hours post-infection, **B)** 18 hours post-infection, **C)** infection with *M. tuberculosis* CDC1551 at 18 hours compared to 4 hours and **D)** infection with *Δppe38-71* at 18 hours compared to 4 hours post infections. Significance cut-offs were set at a q-value less than 0.05 and a log2 fold change greater than 1. Numbers indicate the amount of significantly regulated proteins. **E)** Volcano plot representing regulated proteins of THP-1 macrophage-like cells infected with the complemented strain compared to *M. tuberculosis* CDC1551 at 18 hours post-infection. Significance cut-offs were set at a q-value less than 0.05 and a log2 fold change greater than 1. **F)** Volcano plot representing proteins of THP-1 macrophage-like cells infected with the complemented strain at 18 hours compared to 4 hours post infection. Significance cut-offs were set at a q-value less than 0.05 and a log2 fold change greater than 1. **G)** Heat map of the main regulated proteins between the groups, displaying the $\log_2$ fold changes between THP-1 macrophage-like cells infected with *M. tuberculosis* CDC1551, *Δppe38-71* and the complemented strain at 18 hours post-infection. LFQ intensities were Z-scored and are from three independent biological replicates.

associated with a pro-inflammatory response (Fig. 2C, 2G). In contrast, *M. tuberculosis Δppe38-71*-infected macrophages had a remarkably low pro-inflammatory response to infection relative to wild type-infected macrophages (Fig. 2D, 2G). Macrophages infected with the complemented strain displayed largely a similar temporal profile as those infected with *M. tuberculosis Δppe38-71* (Fig. 2E, 2F). This may be due a number of reasons, the most likely candidates involving the promotor or the physiological state of the bacteria upon infection. As the *hsp60* promotor is constitutively expressed and at high amounts it is unlikely that the lack of complementation in the macrophage is due to a lack of protein products from the *ppe38-71* operon (31). We do indeed observe greater variability within the replicates of the complement strain which likely results in reporting of non-significant results from hypothesis testing (Fig. S1C, S1F). Finally the PE-PGRS proteins are cell surface proteins and the presence of detergent is known to influence the mycobacterial capsule (9,32,33). Thus, the physiological state of the bacteria may be altered due to the presence of detergent in the culture media prior to infection. To gain more insight, we examined the expression of individual proteins and observed at least partial complementation in, among others, interleukin 1 Beta (IL-1B), nuclear factor kappa B (NF-kB) 2 and '-5'-oligoadenylate synthase 2 (OAS2) as indicated by individual fold change values (Fig. 2G, Data S1: Table S2). Lastly, gene set enrichment analysis of the differential response at 18 hours post infection between *M. tuberculosis* CDC1551 and *Δppe38-71*-infected macrophages indicated NF-kB signalling as the most enriched pathway (Fig. S2A). Furthermore, key proteins that were differentially regulated during infection with these strains included IL-1B (Fig. S2B), NF-kB2 Fig. S2C), retinoic acid-inducible gene I (Rig-I/DDX58) (Fig. S2D) and Interferon stimulated gene 15 (ISG15) (Fig. S2E), indicative of both altered interleukin and interferon responses, all of which can be controlled by NF-kB signalling.

While we could not definitively state that the major response observed here is driven by the *ppe38-71* deletion, we do observe distinct trends in certain key proteins associated with an inflammatory response. These observations suggest a role for NF-kB signalling based on enrichment analysis of the significantly regulated proteins, which may shed light on the low inflammatory response of the macrophage to *M. tuberculosis Δppe38-71*. Initiating an inflammatory response is an energy-intensive process and requires upregulation of multiple target proteins and thus increased transcription and translation. To maintain a balance in the proteome, proteins need to be degraded in relation to this increased synthesis (34–36). This phenomenon is known as protein turnover. To further corroborate our findings, the protein turnover of macrophages infected with the different strains, as well as an uninfected control was determined.

## Proteostasis is affected by the presence of PPE38-controlled proteins.

Pulse-chase stable isotope labelling by amino acids in cell culture (pSILAC) can be used to determine protein turnover, and thus proteostasis, of an organism by mass spectrometry (25,37,38). Terminally differentiated THP-1 cells are unable to replicate and thus do not naturally dilute the proteins by division, providing a useful model for studying protein homeostasis. THP-1 macrophage-like cells infected with the *M. tuberculosis* CDC1551, *Δppe38-71* and complemented strains, as well as control uninfected macrophages, were sampled at multiple time points to determine protein turnover in response to infection (Fig. S3A). A total of 1 257 protein half-lives were calculated across all four conditions, after filtering for two unique peptides, the presence of a heavy/light ratio in at least three time points and a coefficient of determination ($R^2$) greater than 0.85 (Data S1 Table S3).

The percentage of heavy amino acid (lysine and arginine) incorporation increased linearly over time to a maximum of 30% incorporation in all conditions (Fig. S3B). Interestingly, no shift in the distribution of protein half-lives occurred between any of the conditions (Fig. S3C). Half-lives were highly correlated within the conditions with coefficients of determination ($R^2$) values found to be 0.92 between *M. tuberculosis Δppe38-71* and *M. tuberculosis* CDC1551-infected macrophages (Fig. S3D), and between macrophages infected with *M. tuberculosis* CDC1551 or the complemented strain (Fig. S3E). The $R^2$ values of half-lives from uninfected macrophages compared to *M. tuberculosis* CDC1551 were only marginally less than that of the uninfected counterparts (Fig. S3F). This was surprising, as it was expected that less correlation would be observed between infected and uninfected macrophages due to increased protein synthesis during infection.

As no overall differences were observed when analysing global protein turnover rates, we used hierarchical clustering to assign proteins with correlated half-lives to distinct clusters. The protein turnover profiles between the uninfected cells and *M. tuberculosis* CDC1551-infected macrophages showed the greatest differences (Fig. 3A). The individual proteins grouped into two distinct clusters, where shifts in the marginal means were primarily driven by *M. tuberculosis* CDC1551 (Fig. 3B, 3C). Deviations of marginal means, the mean half-life of each protein from each group, in response to infection reflect similar trends to those observed in the protein expression profiles. The response to infection with *M. tuberculosis* CDC1551 compared to the uninfected control acts as the driver that separates cluster 1 and 2. While infection by the *M. tuberculosis Δppe38-71* strain has an effect on the protein half-lives compared to uninfected, it is not as pronounced as that observed for wild type. This is in line with our protein ex-

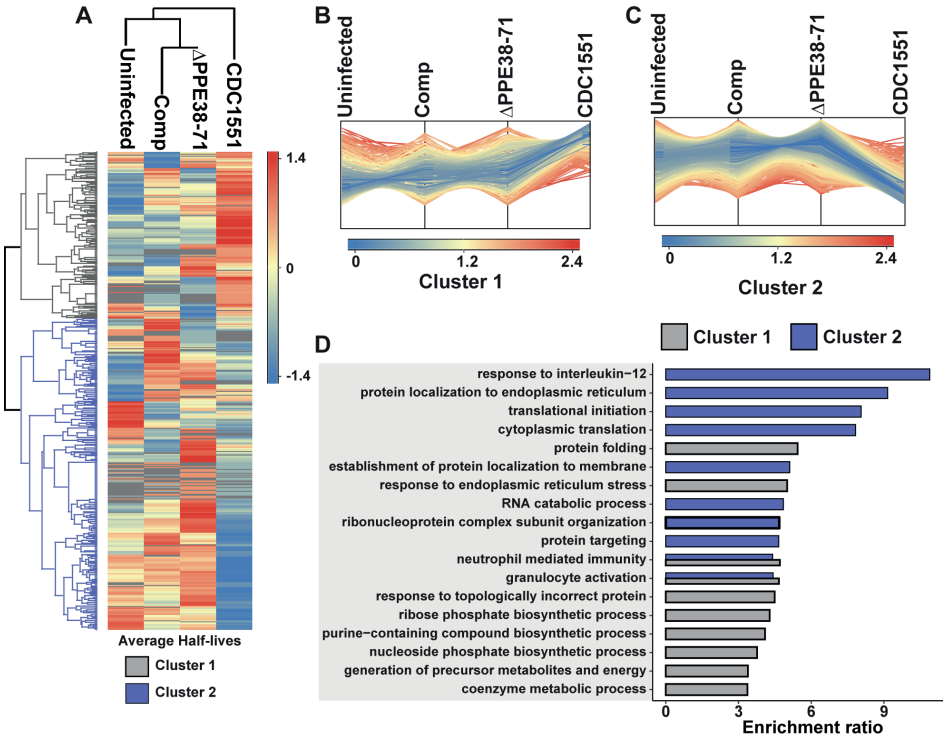



**Figure 3: Infection with *M. tuberculosis* CDC1551 and not *M tuberculosis Δppe38-71* causes a shift in protein turnover associated with the IL-12 pro-inflammatory pathway.**
**A)** Half-lives from THP-1 macrophage-like cells infected with *M. tuberculosis* CDC1551, *M. tuberculosis Δppe38-71* and the complemented strain and uninfected control macrophages were assessed using hierarchical clustering with Euclidean distance. Two distinct clusters formed representing either an increase (cluster 1) or a decrease (cluster 2) in protein half-lives as depicted by profile plots. Protein groups from **B)** cluster 1 and **C)** cluster 2 are represented as a deviation in the estimated marginal mean of each protein half-life, after adjustment of covariates, across each factor. **D)** GO enrichment analysis of each cluster obtained in A and B. Protein turnover data is representative of two biological replicates.

pression results where a modest change was observed during infection. Furthermore, these clusters represent either an increase (cluster 1) or decrease (cluster 2) in protein half-lives and these clusters were used for pathway enrichments. The enrichments indicate that half-lives associated with cluster 1 are involved in proteome integrity and maintenance, while those associated with cluster 2 are involved in the inflammatory response (Fig. 3D). Taken together, macrophages infected with *M. tuberculosis* CDC1551 induce robust inflammatory responses, with dynamic shifts in both expression and half-life. However, these effects are less pronounced when challenged with *M. tuberculosis Δppe38-71*. As in our other proteomics experiments, the complemented strain clustered closer with *M. tuberculosis Δppe38-71* infections than infection by CDC1551 and all infections were distinct from the uninfected profile.

## Protein half-life differences between *M. tuberculosis* CDC1551- and *Δppe38-71*-infected macrophages point to an altered pro-inflammatory response.

To further investigate the effect of protein half-lives of infected macrophages, the differential turnover of individual proteins was assessed using hypothesis testing. The protein half-lives were filtered prior to statistical testing to remove all missing values from the test groups. From these tests, we identified 24 proteins with differentially regulated half-lives between *M. tuberculosis* CDC1551 and *M. tuberculosis Δppe38-71* (Fig. 4A) as well as 10 differentially regulated half-lives between *M. tuberculosis* CDC1551 and macrophages infected with the complemented strain (Fig. 4B). Of these proteins, 17 were unique to macrophages infected with *M. tuberculosis Δppe38-71* compared to *M. tuberculosis* CDC1551-infected macrophages, and three were unique to macrophages infected with the complemented strain compared to *M. tuberculosis* CDC1551-infected macrophages. Seven proteins were shared between these conditions (Fig. 4C). While some proteins are complemented, a similar trend as with the label-free comparisons is observed (Data S1 Table S4).

The proteins displaying differential half-lives in the *M. tuberculosis Δppe38-71* infected compared to *M. tuberculosis* CDC1551-infected macrophages were split into upregulated (slower turnover than in *M. tuberculosis* CDC1551-infected macrophages) and downregulated proteins (faster turnover). These proteins were used for KEGG pathway analysis and showed enrichment of fatty acid degradation pathways in the slower protein turnover cluster. Interestingly, Th1/Th2 cell differentiation was enriched in the faster turnover cluster (Fig. 4D). From the hypothesis test, a number of immune-related proteins displayed differential half-lives. Proteins such as PSMA4, Stat-1 and Tapasin had a rapid turnover in *M. tuberculosis* CDC1551-infected macrophages compared to the other conditions (Fig. S4A). These proteins are involved in MHC class I presentation

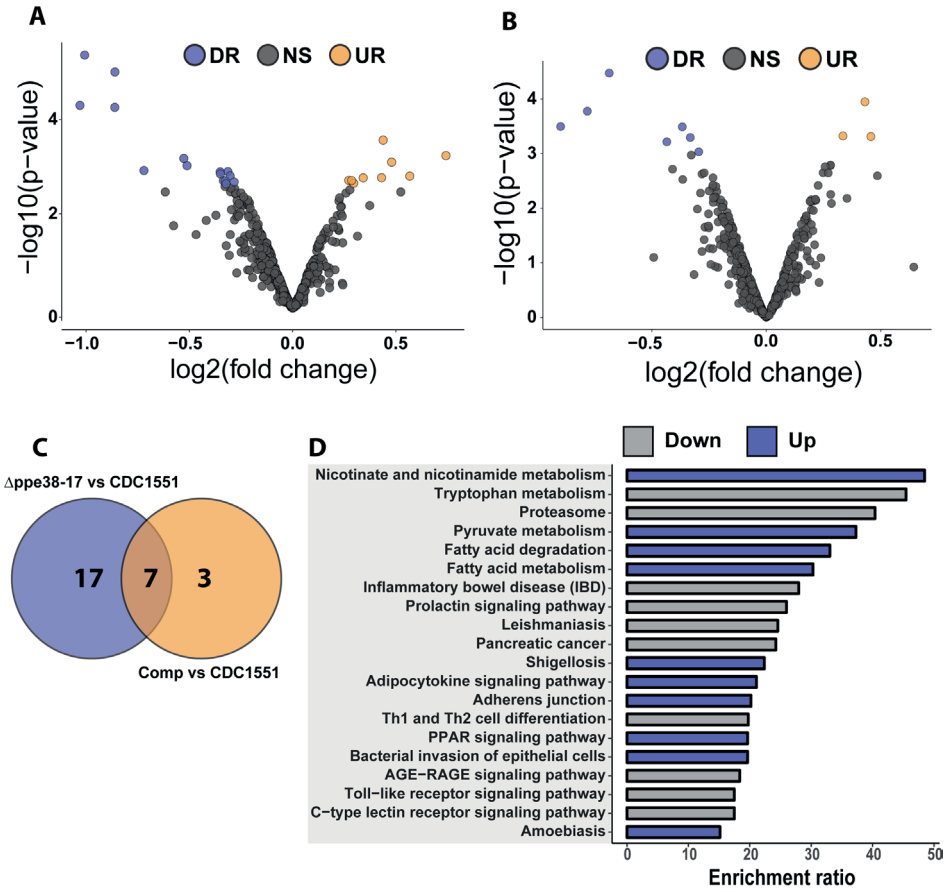



**Figure 4: Differential regulation of individual protein half-lives point towards changes in the inflammatory response.**

Volcano plot depicting altered protein half-lives between **A)** *M. tuberculosis* CDC1551- and *Δppe38-71*-infected macrophages or **B)** *M. tuberculosis* CDC1551- and complement-infected macrophages. Significantly altered protein half-lives are highlighted based on q-value set at 0.05. Downregulated proteins are denoted as DR (purple), upregulated as UR (orange) and non-significant as NS (grey) in the volcano plots. **C)** Venn diagram depicting the number of macrophage protein half-lives present either uniquely or commonly between the *M. tuberculosis* CDC1551 infected and *Δppe38-71* infected macrophage and between *M. tuberculosis* CDC1551 infected and complement-infected macrophage comparisons. Data is representative of two biological replicates. **D)** Over-representation analysis of the proteins with differentially regulated half-lives against KEGG pathways, where up is indicative of rapid turnover and down indicative of slower turnover. Over-representation was calculated using the WebGestalt API using default parameters.

in the case of PSMA4 and Tapasin, or inflammatory cell signalling which is mediated by Stat-1. Interestingly, modulation of the MHC class I antigen presentation pathway by PPE38 has been demonstrated previously (39). Furthermore, proteins affecting innate immunity with slow turnover in *M. tuberculosis* CDC1551-infected macrophages included Glyoxalase I and Leukosialin (Fig. S4B). The glyoxalases are used for the detoxification of α-oxaldehydes, in particular methylglyoxal, in eukaryotic cells (40–42).

Increased methylglyoxal can be produced by excess glucose or lack of phosphates by interfering with the glycolysis pathway and lead to the production of advanced glycation endproducts (AGE) (43). In turn AGE can stimulate the NF-kB pathway resulting in M1 polarization of the macrophage (44). As we see a clear inflammatory response in CDC1551-infected macrophages we expect an increase in the half-life of glyoxalase in order to cope with the detoxification of MG. Likewise, Leukosialin (CD43) is a cell surface protein that has been shown to interact with *M. tuberculosis* Cpn60.2 to induce TNF-α production (45), however, its expression can also commit T-cells to Th1 a response (46,47).

Taken together, there is a robust macrophage response to infection with *M. tuberculosis* CDC1551 while little to no response is observed when challenged with *Δppe38-71*. Mixed results were observed for macrophages infected with the complemented strain thus obscuring the effect and requiring further investigation into the root cause of this differential phenotype. Based on our observations thus far it is certain that infection by the *M. tuberculosis* CDC1551 strain is stimulating the pro-inflammatory pathways of THP-1 macrophage-like cells. This is the expected response to infection by bacteria, however this result is not observed to the same extent when challenged with *M. tuberculosis Δppe38-71*. The loss of PPE38 controlled PE-PGRS proteins eliminate a number of PAMPs which may explain this observation, however the complementation does not fully restore the macrophage responses, which obscures the result.

### *M. tuberculosis Δppe38-71* is only partially complemented *in vitro* and complementation is detergent dependent.

The mixed response observed in the THP-1 macrophage-like cells infected with the complemented strain prompted us to further investigate possible differences between the bacilli under different growth conditions. For this, we analysed the proteome and secretome profiles of the different *M. tuberculosis* strains *in vitro*. As shown above, the PE-PGRS secretion phenotype is complemented at day four of *in vitro* growth (Fig. 1A) and there are no significant differences in optical density at this time point (Fig. 1B). We, therefore, sampled whole-cell lysates and supernatants from four-day-old cultures, which reflects the state of the strains used for infection. As PE-PGRS proteins are associated with the cell surface and thus susceptible to detergents, we removed Tween-80 by sequential washing and allowed all three strains to grow for an additional four days to represent a detergent-free supernatant. We found no clear separation between strains from the whole-cell lysate fraction (Fig. 5A) in the first principal component, thus indicating minimal global variation between strains. However, separation in the first principal component could be observed in supernatant samples from mycobacteria cultured in Tween-80, and the complemented strain clusters with the *Δppe38-71*

strain (Fig. 5B). Interestingly, the secretome profile of the complemented strain clusters closer to *M. tuberculosis* CDC1551 when detergent is omitted from the culture supernatants (Fig. 5C). Western blots of all three fractions were used to determine the localisation and presence of PE-PGRS proteins in these samples. We found the majority of PE-PGRS proteins in the supernatants containing detergent (detergent fraction) and no PE-PGRS proteins in the detergent free supernatants (Fig. 5D).

Comparison of differential protein expression in the whole-cell lysates between *M. tuberculosis* CDC1551 and *M. tuberculosis Δppe38-71* or the complemented strain revealed only two differentially regulated proteins below a q-value cut off set to 0.05, one of which was PPE71 (Data S1 Table S5). However, when comparing the detergent fraction of *M. tuberculosis Δppe38-71* to *M. tuberculosis* CDC1551, downregulation of multiple PE/PPE proteins was observed (Fig. 5E, Data S1 Table S6). Complementation of this phenotype was observed, but partial, as many of the PE/PPE proteins still showed significantly different protein abundance relative to *M. tuberculosis* CDC1551 (Fig. 5F, Data S: Table S6). Finally, comparisons of the detergent-free fraction of *M. tuberculosis Δppe38-71* to *M. tuberculosis* CDC1551 had a total of 142 differentially regulated proteins, 22 of which were downregulated and 120 were upregulated (Fig. S5A, Data S1: Table S7). Complementation in the detergent-free fraction was much more pronounced with 7 proteins significantly upregulated and 2 proteins significantly downregulated as compared to *M. tuberculosis* CDC1551 (Fig. S5B, Data S1: Table S7). However, the PE/PPE proteins did not feature prominently within this fraction and are likely associated with the cell surface.

These results show that the PE-PGRS and PPE-MPTR proteins are indeed expressed in the complemented strain as indicated in Fig. 1A, however to a lesser extent than in wild type. As this represents the state used for infection, it is likely that the lack of full complementation stems from the culturing conditions prior to infection. To corroborate our infection proteomics findings, cell-free supernatants from detergent-free cultures (Fig. 5E and Fig. 5F) were used to stimulate differentiated THP-1 macrophage-like cells. As pro-IL-1B was the most differentially regulated protein, we probed for expression of this protein by Western blot 18 hours after stimulation and found similar differential regulation as observed in our infection proteomics data (Fig. S5C). Macrophages stimulated with *M. tuberculosis* CDC1551 supernatants displayed a significantly higher IL-1B expression compared to *Δppe38-71*, and this phenotype was partially complemented (Fig. S5D, S5E).

By analysing the spatial distribution of proteins in *M. tuberculosis* cultures at their metabolic state before infection, we could show that complementation of PE-PGRS

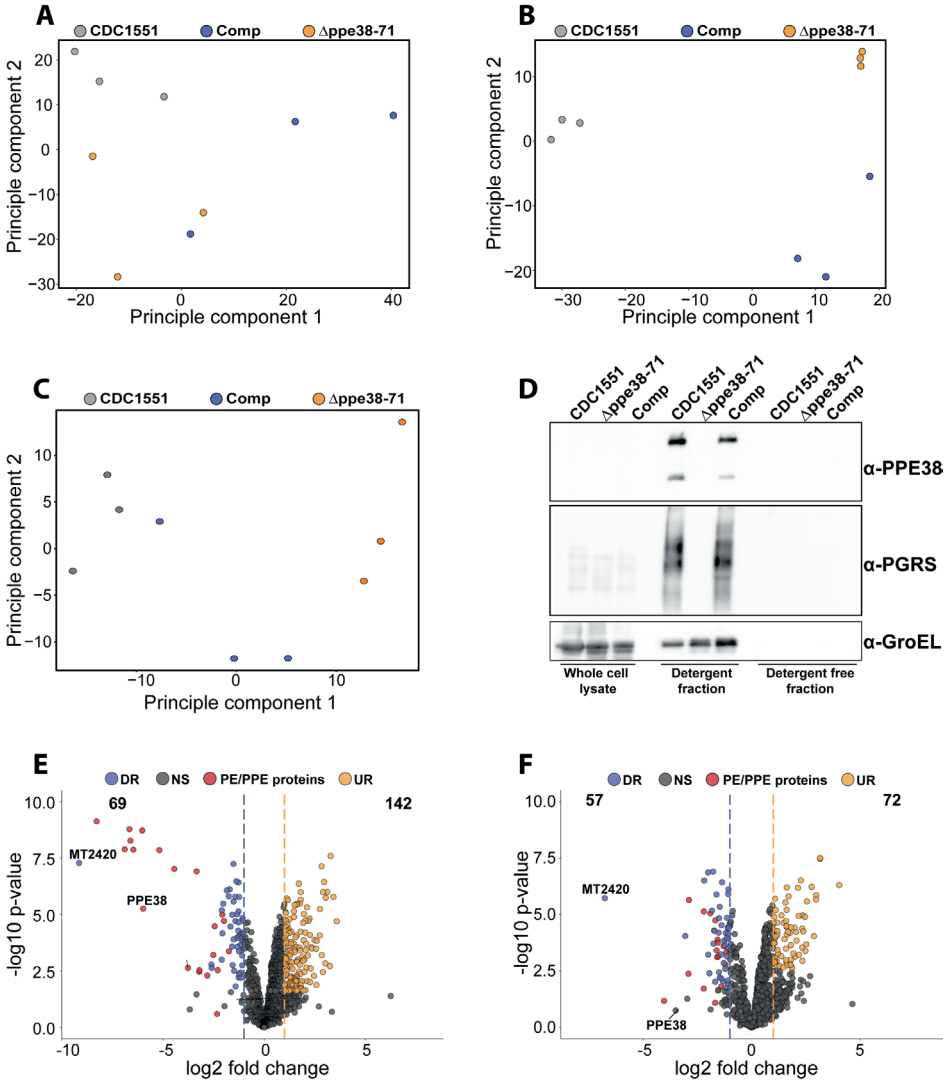



**Figure 5: *In vitro* proteomics of *M. tuberculosis* strains show partial complementation that can be restored by the omission of detergent.**

Principal component analysis of **A)** whole-cell lysates, **B)** supernatant containing detergent and **C)** supernatant without detergent of each *M. tuberculosis* strain after four days of cultivation. **D)** Western blot of each fraction shows the PE-PGRS proteins in the supernatant only when there is detergent present in the growth media. Volcano plots of supernatants containing detergent from **E)** *M. tuberculosis Δppe38-71* and **F)** the complemented strain compared to *M. tuberculosis* CDC1551 when cultivated in detergent. The q-value was set to 0.05 and the log2 fold change was set to 1. The data is representative of three independent experiments.

secretion was not fully restored prior to infection in the complemented strain (Fig. 5E, 5F). However, the PE-PGRS secretion can be restored in the complemented strain when detergent is omitted, which is likely due to the surface localisation of PE/PPE proteins that are detached in the presence of detergent. By stimulating macrophages with cell-free supernatants, representing a high or low abundance of PE/PPE proteins, we could show differential regulation of IL-1B similar to infection with live *M. tuberculosis* CDC1551 and *Δppe38-71* strains. As seen in Fig. 5, the composition of the *M. tuberculosis Δppe38-71* secretome relative to *M. tuberculosis* CDC1551 is a complex mixture of differentially secreted proteins not only limited to PE/PPE proteins. However, the most down-regulated proteins are indeed PE/PPE proteins, with members from both the PE-PGRS and PPE-MPTR sub-families (Fig 5E, Data S1: Table S6). It is thus likely that our observations are driven not by PPE38 alone but rather a physiological state, created by the absence of PPE38, where a group of PE/PPE proteins are acting as the effectors of PPE38 and mediating a response.



## Both RelB and NF-kB p50 are translocated to the nucleus of *M. tuberculosis Δppe38-71* infected macrophages.

NF-kB signalling regulates inflammatory responses in innate immune cells after receptor engagement with PAMPs. NF-kB subunits can be localised in the cytosol as inactive transcription factors or can be activated and translocated to the nucleus (Fig. 6A) (48). We further investigated both NF-kB1 and NF-kB2, representing canonical and non-canonical NF-kB signalling pathways, respectively. The canonical NF-kB1 p105 and p50 subunits were detected in all tested conditions (Fig. 6B, NF-kB1). Little to no cleavage of NF-kB2 was observed at 18 hours post-infection in any of the conditions (Fig. 6B, NF-kB2 p52). Upregulation of the p100 subunit and pro-IL-1B (Fig. 6B, NF-kB2 p52; IL-1B) was observed for *M. tuberculosis* CDC1551-infected macrophages. This is congruent with our other observations and indicative of a pro-inflammatory response in CDC1551-infected macrophages (49). Furthermore, our proteomics results suggest an increased interferon signalling in *M. tuberculosis* CDC1551-infected macrophages. We used ISG15 as our downstream target to validate these results and indeed found upregulation of ISG15 in *M. tuberculosis* CDC1551-infected macrophages at 18 hours compared to other conditions. We did not directly observe the STING pathway in the proteomics results, but did observe evidence for RIG-I induction. However, the interferon pathway is reportedly stimulated by STING in macrophages infected with *M. tuberculosis* (50). We therefore used Western blotting to validate expression levels of this receptor and found no differences in RIG-I expression in any of the infection conditions (Fig. S6). This suggests that the RIG-I is not responsible for the differential ISG15 expression, which is in line with other observations (50).

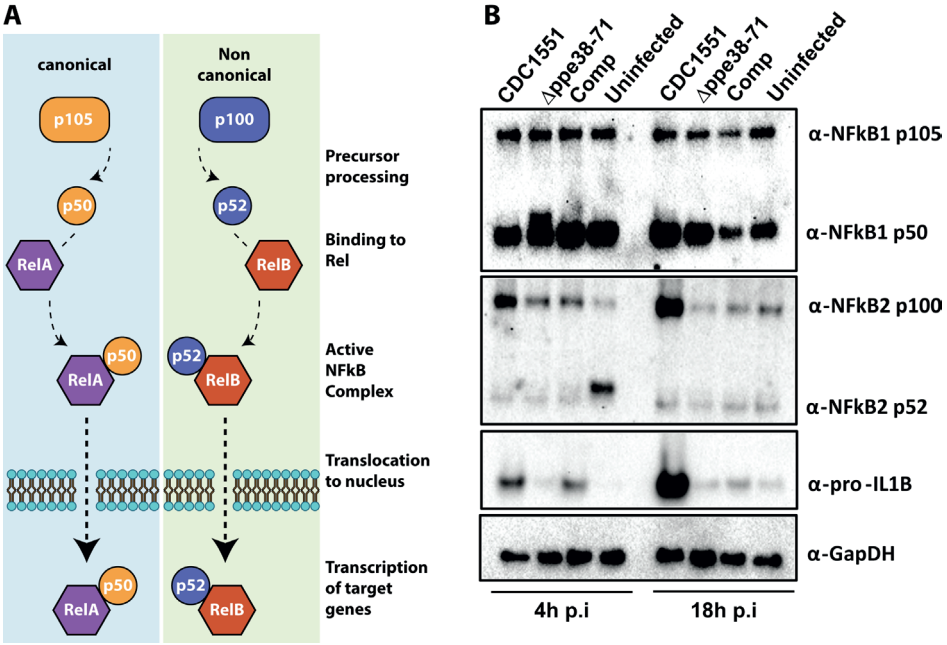



**Figure 6: THP-1 macrophage-like cells signal through the canonical NF-kB pathway when infected with *M. tuberculosis* CDC1551.**

**A)** Schematic representation of the canonical and non-canonical NF-kB pathways. **B)** Western blots for components of the NF-kB pathway (α-NF-kB1 and α-NFkB2) and downstream effectors (α-pro-IL1β) as well as their respective sub-units. The Western blots were done on protein samples harvested at 4 hours and 18 hours post-infection. α-GapDH was used as a loading control.

Based on these results, *M. tuberculosis* CDC1551 is triggering the canonical inflammatory NF-kB pathway. Activation of NF-kB1 and/or NF-kB2 results in the translocation of either RelA or RelB, respectively. These Rel proteins act as the effectors of this pathway and initiate transcription of target genes, depending on the required immune response. Previous studies have shown increased NF-kB1 p50 levels in the nucleus of murine macrophages infected with wild type *M. marinum* compared to a *ppe38* transposon mutant (17). While a different technique and model organism was used, this result is in line with our observations. However, it remains unclear whether *M. tuberculosis Δppe38*-71 reduces or entirely inhibits canonical NF-kB signalling or whether it induces an alternative signalling pathway.

To further investigate the activation of NF-kB signalling in infected macrophages, we used confocal microscopy to determine RelA and RelB nuclear translocation in infected macrophages. RelA translocation was observed in macrophages infected with *M. tuberculosis* CDC1551 and in macrophage infected with the complemented strain after 18 hours of infection (Fig. 7A, Fig. 7C). However, in contrast, no RelA translocation was

observed for macrophages infected with the *Δppe38-71* strain (Fig. 7B) or the uninfected control (Fig. S7A). This result was surprising as it was expected that the inflammatory response would be triggered by RelA translocation, although perhaps to a limited extent as compared to *M. tuberculosis* CDC1551-infected macrophages. The absence of RelA translocation in the *M. tuberculosis Δppe38-71* infected macrophages prompted us to investigate RelB translocation. While no RelB translocation was observed for *M. tuberculosis* CDC1551- infected macrophages or uninfected control (Fig. 7D, Fig. S7B), RelB did indeed translocate to the nucleus in *M. tuberculosis Δppe38-71* infected macrophages (Fig. 7E). We also observed translocation of RelB in the complemented strain-infected macrophages (Fig. 7F), however, to a lesser extent than in *M. tuberculosis Δppe38-71* infected macrophages (Fig. 7J). The duality of the translocation events observed for the



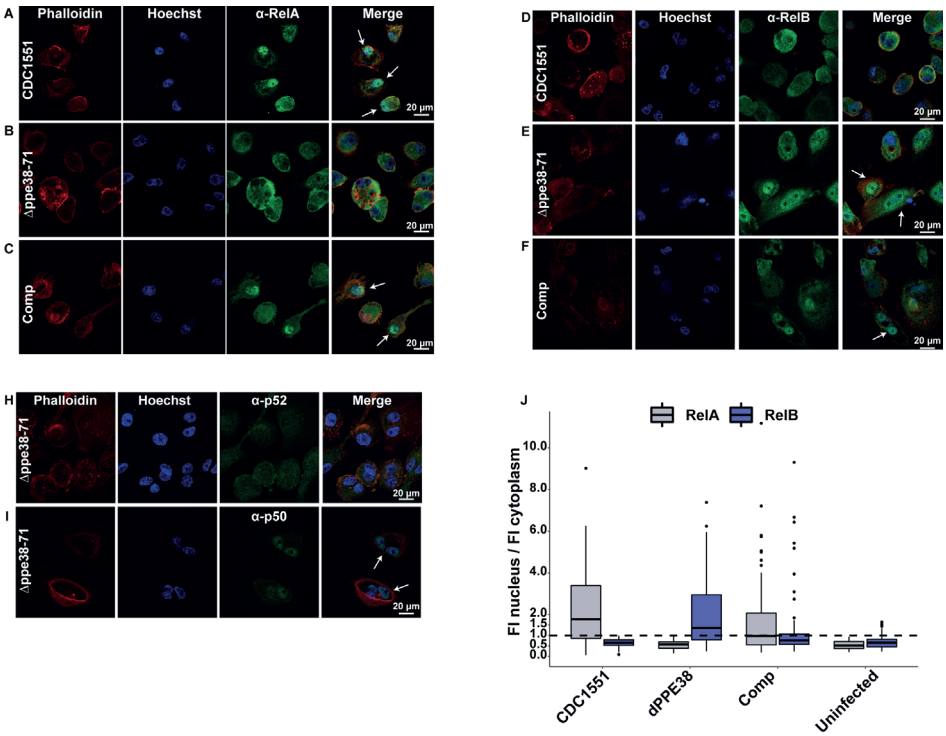



**Figure 7: Infection with *M. tuberculosis Δppe38-71* stimulates a RelB/p50 pathway in THP-1 macrophage-like cells.**

Confocal microscopy images showing individual fluorescent channels probing for F-actin (phalloidin), the nucleus (Hoechst) and **A-C)** RelA (Alexa Fluor 488) for **A)** *M. tuberculosis* CDC1551, **B)** *Δppe38-71,* **C)** complement-infected macrophages. Macrophages infected with **D)** *M. tuberculosis* CDC1551, **E)** *Δppe38-71* and **F)** the complemented strain were also probed for translocation of RelB (Alexafluor 488). **H)** Confocal microscopy images showing p100/p52 and **I)** p105/p50 nuclear translation in *Δppe38-71* infected macrophages. White arrows indicate translocation events. **J)** Quantification of fluorescence intensity in the nucleus compared to that in the cytoplasm. Twenty cells were chosen from at least 10 random fields per replicate with a minimum of three independent experiments. The dashed line indicates a ratio of 1 for the cut off where translocation occurs.

complemented strain was an interesting observation, which could give an additional explanation as to why partial restoration of the inflammatory phenotype is observed. Macrophages infected with the complemented strain seemingly mimic both the wild type, in RelA translocation, and the *Δppe38-71* phenotype, in RelB translocation. This, in turn, may result in differential cytokine release on a per-cell basis. Finally, low transient levels of RelB translocation was observed in macrophages infected with the complemented strain (Fig. 7J). This phenomenon is consistent with other observations and likely due to poor binding of IkB-α allowing for transient movement to the nucleus for RelB complexes (51–53).

Translocation of RelB suggests that *M. tuberculosis Δppe38-71* stimulates the non-canonical NF-kB pathway, which can cause a slow, yet persistent, inflammatory response instead of a burst response (54). However, we did not observe upregulation of the p52 non-canonical partner in *M. tuberculosis Δppe38-71* infected macrophages. Furthermore, slow and persistent inflammation should accumulate IL-1B over time, yet a downregulation of this cytokine was observed. Nonetheless, the p50 subunit was available as a potential partner for translocation. Based on our observations in both mass spectrometry, Western blots and microscopy, macrophages infected with *M. tuberculosis* CDC1551 are signalling through the canonical pathway characterised by RelA and p50 translocation. The translocation dynamics of macrophages infected with *M. tuberculosis Δppe38-71* strain is however unclear. To rule out the non-canonical pathway, macrophages were infected with *M. tuberculosis Δppe38-71* and probed for translocation of p50 and p52 at 18 hours post-infection. No translocation of the non-canonical NF-kB2 p52 subunit was observed (Fig. 7H), while the canonical NF-kB1 p50 subunit was found to localise to the nucleus (Fig. 7I). No noticeable translocation of either p50 or p52 was observed in the uninfected THP-1 macrophage-like cells (Fig S7C, Fig. S7D). This indicates that infection with *M. tuberculosis Δppe38-71* causes translocation of RelB/p50 complexes (Fig. S7E), which has been shown to be associated with anti-inflammatory responses (55–57).

## Less IL-12p70 is found in supernatants of *M. tuberculosis Δppe38-71* infected macrophages at 48h post-infection.

Pro- and anti-inflammatory macrophage responses are characterised by the secretion of specific cytokines. Examples include the secretion of IL-12p70 as a marker for the pro-inflammatory response and IL-13 to represent the anti-inflammatory response (58). Macrophages were incubated for 48 hours to examine the delayed inflammatory response during infection with *M. tuberculosis Δppe38-71*. We found a greater abundance of IL-12p70 in *M. tuberculosis* CDC1551 and complement-infected macrophages relative to the *M. tuberculosis Δppe38-71* infected, and uninfected control macrophages

(Fig. 8A). This differential secretion pattern indicates that 48 hours post-infection the macrophages have still not launched a strong inflammatory response to the *M. tuberculosis Δppe38-71* strain as compared to *M. tuberculosis* CDC1551. There was also no significant difference in the secretion of IL-12p70 in *M. tuberculosis Δppe38-71* infected macrophages when compared to the uninfected control (Fig. 8A, Data SI: Table S8). In contrast to this, IL-13 displayed the opposite pattern with less IL-13 present in the supernatant of *M. tuberculosis* CDC1551 and complement-infected macrophages compared to macrophages infected with *M. tuberculosis Δppe38-71* (Fig. 8B).

Taken together, these results indicate that the PE/PPE proteins controlled by PPE38 have an effect on modulating macrophage responses through NF-kB signalling. THP-1 macrophage-like cells infected with *M. tuberculosis* CDC1551 are exposed to substantially more PAMPs as seen in Fig. 5E and thus a canonical RelA/NF-kB1 p50 pathway is initiated resulting in a strong pro-inflammatory response (Fig. 8C). However, in the absence of PPE38 and its effectors, RelB/p50 translocates to the nucleus and likely dampens this response, resulting in an anti-inflammatory phenotype over time (Fig. 8D).



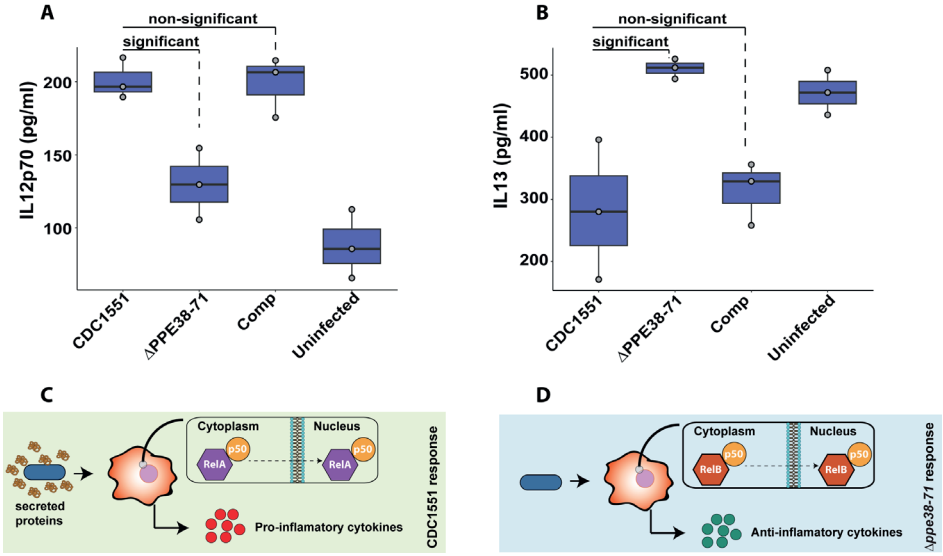



**Figure 8: Interleukin-12p70 and IL-13 are differentially secreted by THP-1 macrophage-like cells at 48h post-infection.** The levels of IL-12p70 **A)** and IL-13 **B)** in supernatants of THP-1 macrophage-like cells infected with either *M. tuberculosis* CDC1551, *Δppe38-71* or complemented strains and uninfected macrophages were determined by ELISA. Data is representative of three biological replicates and significant differences were determined with one-way ANOVA as well as a Tukey HSD post-hoc test with a q-value set at 0.05. Schematic illustration of a proposed macrophage response upon infection with **C)** *M. tuberculosis* CDC1551 and **D)** *M. tuberculosis Δppe38-71*.

## DISCUSSION

*M. tuberculosis* can influence disease outcome by altering protective host innate immune responses. However, the effector molecules involved in host-pathogen interactions and altered host responses remain ill-defined. It was previously demonstrated that deletion of *ppe38-71* in *M. tuberculosis* resulted in increased virulence in a murine infection model (13). In this study, we investigated the role of *M. tuberculosis Δppe38-71* in the context of host-pathogen interactions to identify the mechanisms associated with increased virulence. An important observation made here is that the PE-PGRS and PPE-MPTR proteins controlled by PPE38 are responsible for driving altered macrophage responses. These proteins are unique to pathogenic mycobacteria and several authors have proposed a role for PE-PGRS and PPE-MPTR proteins in host-pathogen interactions (7).

Infection of macrophages with wild type *M. tuberculosis* CDC1551 results in increased expression of pro-inflammatory cytokines and chemokines, specifically, upregulation of pro-IL-1B, several interferon-inducible proteins (ISG15, IFIT1, IFIT2, IFIT3, IFIH1 and MX1) and rapid recycling of components associated with IL-12 signalling. Conversely, deletion of *ppe38-71* in *M. tuberculosis* resulted in a dampened innate immune response in infected macrophages, where IL-1B was found to be downregulated over time. This shift away from a pro-inflammatory response was observed in both temporal label-free proteomic analysis as well as protein turnover analysis using pulse SILAC labelling. However, there seems to be a disparity in the cytokine signalling of dendritic cells and macrophages when infected with ppe38 mutants in the current literature. After overnight stimulation of murine dendritic cells with *M. tuberculosis* CDC1551, *M. tuberculosis Δppe38-71* and the complemented strain there was no significant difference observed in supernatant IL-12p40/70, IL-6 and TNF-α (18). In contrast, there is a stark difference in IL-6 and TNF-α levels found in macrophages infected with the *M. marinum* wild type and a *ppe38* transposon mutant strain (16). In the latter case, TNF-α is lower in macrophages infected with a *M. marinum ppe38* transposon mutant compared to wild type at both 24 and 48 hours post infection. We did not assay for IL-6 or TNF-α, however from our cytokine assay we find less IL-12p70 in the supernatant of macrophages infected with *M. tuberculosis Δppe38-71*. Notably the difference in abundance of IL-12p70 detected in this study could easily be influenced by various factors such as cell type, MOI and timeframe in comparison to the levels measured in dendritic cells (18).

Macrophages challenged with concentrated supernatants from each strain showed the same regulation of IL-1B as during live mycobacterial infections. Here it was demonstrated that the deletion of the *ppe38-71* operon has farther reaching consequences

to the cell than only this protein or even the PE-PGRS proteins. It is thus likely not one single protein is driving the macrophage response but a conglomerate of proteins representing a physiological state upon deletion of the *ppe38-71* operon. Many of the PE/PPE proteins have unknown functions, however, multiple studies have shown that these proteins are upregulated during macrophage infection (59–61) and are highly immunogenic (62). In Fig. 5E several PE/PPE proteins were identified as being controlled by PPE38 and these are the likely candidates for intracellular effector proteins. Notably, PPE10 was represented in this cluster and has been implicated in the disruption of *M. marinum* capsule integrity, altering colony morphology and attenuation of virulence in zebrafish (63). As PPE10 was one of the most downregulated proteins in the *M. tuberculosis Δppe38-71* strain, it is likely similar phenotypes can be expected from PPE38 knock outs. However, no morphology differences were visible in *M. tuberculosis Δppe38-71,* although the upregulated proteins in Fig. 5E were associated with intracellular proteins. While not definitive, this gives some indication of a damaged cell wall. Interestingly, *M. marinum ppe38* transposon mutants demonstrate a visible change in colony morphology (16). It is as of yet unclear whether this is due to differential regulation of PPE10 by proxy of PPE38. Nevertheless, drawing this conclusion seems likely given the evidence. Taken together, the *in vitro* profile of *M. tuberculosis Δppe38-71* prior to infection indicates decreased levels of multiple virulence factors known to induce pro-inflammatory responses in macrophages. Based on these results it is likely not one protein that drives the effect but many proteins that are altered due to the loss of the *ppe38-71* operon.



Functional enrichments revealed altered NF-kB signalling between macrophages infected with *M. tuberculosis* CDC1551 and those infected with *M. tuberculosis Δppe38-71*. Canonical NF-kB signalling is partly responsible for the induction of pro-inflammatory responses and is characterised by the translocation of the RelA protein along with the NF-kB1 subunit in a RelA/p50 complex (49). This canonical signalling pathway is induced during infection with *M. tuberculosis* CDC1551. In contrast, infection by *M. tuberculosis Δppe38-71* stimulated the signalling of a different NF-kB pathway where RelB and p50 are translocated to the nucleus. A multi-organ inflammatory response observed in mice with a $RelB^{(-/-)}$ knock out was aggravated in a $RelB^{(-/-)}/p50^{(-/-)}$ double knock out (55,64). This phenotype indicates that inflammation is controlled by the RelB/p50 pathway and is likely used to limit excessive inflammation during activation of the canonical NF-kB pathway. Furthermore, a study investigating responses in dendritic cells and macrophages stimulated with LPS has shown that the RelB/p50 pathway inhibits TNF-α production by modulating the canonical pathway (57). We have shown that in the absence of PPE38, this pathway is activated in infected THP-1 macrophage-like cells, which provides a molecular mechanism that could be used by *M. tuberculosis* to

drive switching of inflammatory states in macrophages during infection. In addition, differential localisation of NF-kB subunits has also been previously reported for an *M. marinum ppe38* transposon mutant as revealed by spatial proteomics (16,17).

Based on the results reported here and by others, the secretion of PPE38-dependent proteins to the extracellular milieu, where these proteins are able to interact with host proteins, can initiate a differential inflammatory cascade. In the absence of these effectors, immune dampening is observed mediated by RelB as the likely molecular switch. The *ppe38-71* region is indeed a hotspot for evolutionary activity which includes recombination events, truncations, gene fusion formation and more recently a source for phenotype sharing as a donor for horizontal gene transfer (65–68). Thus, a natural deletion of the *ppe38-71* operon can confer an evolutionary advantage by dampening the innate immune response and possibly by providing a downstream molecular mechanism for controlling macrophage polarisation states. A bacterial mechanism to dampen macrophage responses and switch the polarisation state has been shown to be mediated by effectors of the Spi-2 secretion system in Salmonella (69). Early investigations into *M. tuberculosis* HN878, a member of the lineage 2 isolates of *M. tuberculosis*, demonstrated increased virulence associated with the failure to stimulate Th1 responses (70), similar to the observations made in this study. Interestingly, the same study found that a lack of a pro-inflammatory response was associated with an increased induction of type I interferons (70). In this study we observe increased ISG15 expression, which we speculate may be as a result of the induction of the STING pathway (50). The increased production of ISG15 may partly or wholly be caused by an increase in type I interferons elicited by the *ppe38-71* mutant (71). It was further demonstrated that the decreased inflammatory response was associated with the presence of a phenolic glycolipid on the cell surface of *M. tuberculosis* HN878 (72). This phenolic glycolipid is synthesised by an intact copy of the *pks15/1* gene found in a subset of lineage two isolates (72). Later studies investigated whether an intact *pks15/1* gene confers the same hypervirulence, low inflammatory response phenotype regardless of the genetic background. The authors found that the phenolic glycolipid can act to modulate the host cytokine response but does not directly extend to a hypervirulent phenotype (73). The authors further speculate that the phenolic glycolipid forms a part of a greater genotypic and phenotypic profile of the lineage 2 strains to confer the dampened immune response and hypervirulence (73). Interestingly, we have previously found that the *ppe38-71* deletion occurs overwhelmingly within the lineage two isolates (13,68). Furthermore, other studies have demonstrated an increase in virulence of lineage 2 isolates with a naturally occurring *ppe38-71* deletion, this virulence was partially mitigated by the heterologous introduction of this operon (13). Taken together, it is likely that the *ppe38-71* mutation, in part, plays a role in the increased virulence associated

with the lineage two isolates of *M. tuberculosis* by inducing a more permissive environment for bacterial growth during infection. This is supported by the observation that a *ppe38-71* deletion mutant showed increased bacterial load at later stages of infection in mice compared to the wild type parental strain (13). Interestingly, a recent study demonstrated a similar response in IL-1B modulation, where clinical isolates that induce lower levels of IL-1B were able to successfully evade the macrophages response through decreased inflammasome activation (74). Furthermore, isolates that were associated with severe tuberculosis in patients presented with lower cytokine responses in infected peripheral blood monocytes (74). This shows that *M. tuberculosis* is capable of modulating the inflammatory response through multiple molecular mechanisms and does so by selecting for genomic variation that results in decreased inflammatory responses and increased pathogenicity.



In conclusion, we have used complementary mass spectrometry-based approaches along with follow-up validation to elucidate the role of PPE38-controlled proteins in host-pathogen interactions. Wild type *M. tuberculosis* CDC1551 strains induced the canonical NF-kB pathway to stimulate pro-inflammatory responses in infected human macrophages, whereas in the absence of PPE38-controlled PE-PGRS and PPE-MPTR proteins the alternative RelB/p50 NF-kB pathway is induced. This results in an anti-inflammatory phenotype where the macrophages fail to launch an appreciable pro-inflammatory response. Future experiments will have to identify which PE-PGRS and/or PPE-MPTR protein plays a key role in the RelB-mediated switch between macrophage polarisation states that can influence the infectious process.

## DATA AVAILABILITY

All raw mass spectrometry data was deposited to the ProteomeXchange consortium via the PRIDE partner repository under the following accessions: *Mycobacterium tuberculosis* whole-cell lysates (PXD020814), *Mycobacterium tuberculosis* Tween-80 containing secretomes (PXD020813), *Mycobacterium tuberculosis* Tween-free secretomes (PXD021168), THP-1 label-free whole-cell lysates (PXD021167) and THP-1 SILAC labelled whole-cell lysates (PXD021166).

Excel macro used for calculating protein half-lives is available at the following url: https://newtonexcelbach.wordpress.com/downloads download Linest-poly.xls.

The R script and data files used to analyse protein turnover data can be found at the following GitHub repository: https://github.com/JamesGallant/Protein_turnover

## ACKNOWLEDGMENTS

JG would like to acknowledge the NRF for financial support under the NRF-VU Desmond Tutu Doctoral training program and the Harry Crossley Foundation for project support.

TH was supported by a South African National Research Foundation-Department of Science and Technology Innovation Postdoctoral Fellowship (SFP13071721852).

SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NRF.

The authors acknowledge the SA MRC Centre for TB Research and DST/NRF Centre of Excellence for Biomedical Tuberculosis Research for financial support for this work.

## AUTHOR CONTRIBUTIONS

**JG:** conceptualisation, formal analysis, investigation, data curation, visualization, software, writing, funding acquisition, project administration

**TH**: conceptualisation, data curation, methodology, funding acquisition, supervision, writing – review and editing.

**CB**: investigation, writing - review & editing.

**KS**: investigation, writing - review & editing

**SB**: investigation

**IM**: methodology, resources, writing - review & editing

**WB**: conceptualisation, data curation, resources, supervision, project administration, funding acquisition, writing - review & editing

**SS**: conceptualisation, data curation, resources, supervision, project administration, funding acquisition, writing - review & editing

## SUPPLEMENTARY DATA

Supplementary data for not in this document can be accessed at the following URL with a Mendeley account:

https://tinyurl.com/5u3sumvm

## REFERENCES


1. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci [Internet]. 2012 Mar 19 [cited 2019 Mar 7];367(1590):850–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22312052
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jage BB, Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature [Internet]. 1998 Jun 11 [cited 2019 Mar 11];393(6685):537–44. Available from: http://dx.doi.org/10.1038/31159
3. Abdallah AM, Verboom T, Weerdenburg EM, Gey van Pittius NC, Mahasha PW, Jiménez C, et al. PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. Mol Microbiol [Internet]. 2009 Aug 1 [cited 2019 Apr 18];73(3):329–40. Available from: http://doi.wiley.com/10.1111/j.1365-2958.2009.06783.x
4. Houben ENGG, Korotkov K V., Bitter W. Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta [Internet]. 2014 Aug;1843(8):1707–16. Available from: http://dx.doi.org/10.1016/j.bbamcr.2013.11.003
5. Gey Van Pittius NC, Gamieldien J, Hide W, Brown GD, Siezen RJ, Beyers AD. The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G+C Gram-positive bacteria. Genome Biol [Internet]. 2001 [cited 2017 Oct 12];2(10):RESEARCH0044. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11597336
6. Poulet S, Cole ST. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in Mycobacterium tuberculosis. Arch Microbiol [Internet]. 1995 Feb [cited 2019 Mar 11];163(2):87–95. Available from: http://link.springer.com/10.1007/BF00381781
7. Sampson SL. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol [Internet]. 2011;2011(Figure 1):497203. Available from: http://dx.doi.org/10.1155/2011/497203
8. Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, et al. Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol Microbiol [Internet]. 2012 Mar [cited 2019 Mar 11];83(6):1195–209. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22340629
9. Brennan MJ, Delogu G, Chen Y, Bardarov S, Kriakov J, Alavi M, et al. Evidence that Mycobacterial PE_PGRS Proteins Are Cell Surface Constituents That Influence Interactions with Other Cells. Infect Immun [Internet]. 2001 Dec 1 [cited 2019 Mar 7];69(12):7326–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11705904
10. Ates LS, Houben ENG, Bitter W. Type VII Secretion: A Highly Versatile Secretion System. Microbiol Spectr [Internet]. 2016 Feb [cited 2019 Mar 7];4(1):9–19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26999398
11. Banu S, Honore N, Saint-Joanis B, Philpott D, Prevost MC, Cole ST. Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens? Mol Microbiol [Internet]. 2002;44(1):9–19. Available from: http://dx.doi.org/10.1046/j.1365-2958.2002.02813.x
12. Burggraaf MJ, Speer A, Meijers AS, Ummels R, Van Der Sar AM, Korotkov K V., et al. Type VII secretion substrates of pathogenic mycobacteria are processed by a surface protease. MBio. 2019 Oct 29;10(5):e01951-19.

13. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nat Microbiol [Internet]. 2018 Feb 15 [cited 2018 Mar 6];3(2):181–8. Available from: http://www.nature.com/articles/s41564-017-0090-6
14. Ates LS, Dippenaar A, Sayes F, Pawlik A, Bouchier C, Ma L, et al. Unexpected Genomic and Phenotypic Diversity of Mycobacterium africanum Lineage 5 Affects Drug Resistance, Protein Secretion, and Immunogenicity. Genome Biol Evol [Internet]. 2018 Aug 1 [cited 2021 May 24];10(8):1858–74. Available from: https://pubmed.ncbi.nlm.nih.gov/30010947/
15. Orgeur M, Frigui W, Pawlik A, Clark S, Williams A, Ates LS, et al. Pathogenomic analyses of mycobacterium microti, an esx-1-deleted member of the mycobacterium tuberculosis complex causing disease in various hosts. Microb Genomics [Internet]. 2021 Feb 2 [cited 2021 May 24];7(2):1–18. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000505
16. Dong D, Wang D, Li M, Wang H, Yu J, Wang C, et al. PPE38 modulates the innate immune response and is required for Mycobacterium marinum virulence. Infect Immun [Internet]. 2012 Jan [cited 2019 Mar 7];80(1):43–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22038915
17. Wang H, Dong D, Tang S, Chen X, Gao Q. PPE38 of Mycobacterium marinum triggers the cross-talk of multiple pathways involved in the host response, as revealed by subcellular quantitative proteomics. J Proteome Res [Internet]. 2013 May 3 [cited 2019 Mar 7];12(5):2055–66. Available from: http://dx.doi.org/10.1021/pr301017e
18. Ates LS, Sayes F, Frigui W, Ummels R, Damen MPM, Bottai D, et al. RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. PLoS Pathog [Internet]. 2018 [cited 2019 Mar 7];14(6):e1007139. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29912964
19. Ramagli LS, Rodriguez L V. Quantitation of microgram amounts of protein in two-dimensional polyacrylamide gel electrophoresis sample buffer. Electrophoresis [Internet]. 1985 Jan 1 [cited 2019 Mar 26];6(11):559–63. Available from: http://doi.wiley.com/10.1002/elps.1150061109
20. Lu X, Zhu H. Tube-Gel Digestion. Mol Cell Proteomics [Internet]. 2005 Dec [cited 2019 Mar 7];4(12):1948–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16150870
21. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
22. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V., Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805.
23. Gallant JL, Heunis T, Sampson SL, Bitter W. ProVision: A web based platform for rapid analysis of proteomics data processed by MaxQuant. Bioinformatics [Internet]. 2020 Jul 8;36(19):4965–7. Available from: https://doi.org/10.1093/bioinformatics/btaa620
24. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics [Internet]. 2014 Sep [cited 2019 Oct 13];13(9):2513–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24942700

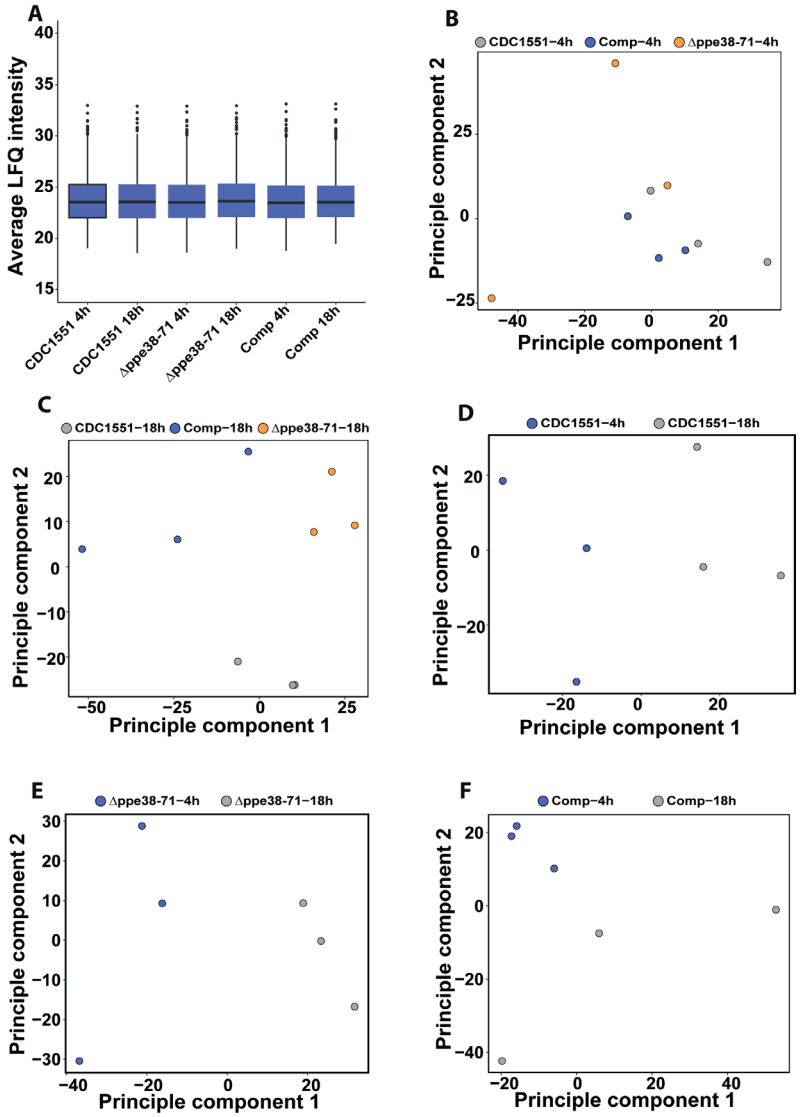
25. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature [Internet]. 2011 May 19 [cited 2019 Mar 7];473(7347):337–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21593866
26. Visscher M, De Henau S, Wildschut MHE, van Es RM, Dhondt I, Michels H, et al. Proteome-wide Changes in Protein Turnover Rates in C. elegans Models of Longevity and Age-Related Disease. Cell Rep [Internet]. 2016 Sep 13;16(11):3041–51. Available from: https://doi.org/10.1016/j.celrep.2016.08.025
27. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res [Internet]. 2017 Jul 3 [cited 2019 Mar 7];45(W1):W130–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx356
28. Iantomasi R, Sali M, Cascioferro A, Palucci I, Zumbo A, Soldini S, et al. PE_PGRS30 is required for the full virulence of Mycobacterium tuberculosis. Cell Microbiol [Internet]. 2012 Mar 1 [cited 2019 Oct 13];14(3):356–67. Available from: http://doi.wiley.com/10.1111/j.1462-5822.2011.01721.x
29. Saini NK, Baena A, Ng TW, Venkataswamy MM, Kennedy SC, Kunnath-Velayudhan S, et al. Suppression of autophagy and antigen presentation by Mycobacterium tuberculosis PE_PGRS47. Nat Microbiol [Internet]. 2016 Sep 15 [cited 2019 Oct 13];1(9):16133. Available from: http://www.nature.com/articles/nmicrobiol2016133
30. Bansal K, Sinha AY, Ghorpade DS, Togarsimalemath SK, Patil SA, Kaveri S V., et al. Src homology 3-interacting domain of Rv1917c of Mycobacterium tuberculosis induces selective maturation of human dendritic cells by regulating PI3K-MAPK-NF-κB signaling and drives Th2 immune responses. J Biol Chem. 2010 Nov 19;285(47):36511–22.
31. Stover CK, De La Cruz VF, Fuerst TR, Burlein JE, Benson LA, Bennett LT, et al. New use of BCG for recombinant vaccines. Nature [Internet]. 1991 [cited 2021 Mar 5];351(6326):456–60. Available from: https://www.nature.com/articles/351456a0
32. Sani M, Houben ENG, Geurtsen J, Pierson J, de Punder K, van Zon M, et al. Direct Visualization by Cryo-EM of the Mycobacterial Capsular Layer: A Labile Structure Containing ESX-1-Secreted Proteins. Ramakrishnan L, editor. PLoS Pathog [Internet]. 2010 Mar 5 [cited 2021 Mar 5];6(3):e1000794. Available from: https://dx.plos.org/10.1371/journal.ppat.1000794
33. Kalscheuer R, Palacios A, Anso I, Cifuente J, Anguita J, Jacobs WR, et al. The Mycobacterium tuberculosis capsule: A cell structure with key implications in pathogenesis [Internet]. Vol. 476, Biochemical Journal. Portland Press Ltd; 2019 [cited 2021 Mar 5]. p. 1995–2016. Available from: /pmc/articles/PMC6698057/
34. Lin J, Ficht TA. Protein synthesis in Brucella abortus induced during macrophage infection. Infect Immun [Internet]. 1995 Apr 1 [cited 2019 Mar 14];63(4):1409–14. Available from: https://iai.asm.org/content/63/4/1409.long
35. Beisel WR. Magnitude of the host nutritional responses to infection. Am J Clin Nutr [Internet]. 1977 Aug 1 [cited 2019 Mar 18];30(8):1236–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/407784
36. Volpe E, Cappelli G, Grassi M, Martino A, Serafino A, Colizzi V, et al. Gene expression profiling of human macrophages at late time of infection with Mycobacterium tuberculosis. Immunology [Internet]. 2006 Jul 10 [cited 2019 Mar 18];0(0):060710044926002-??? Available from: http://doi.wiley.com/10.1111/j.1365-2567.2006.02378.x

37. Lam YW, Lamond AI, Mann M, Andersen JS. Analysis of Nucleolar Protein Dynamics Reveals the Nuclear Degradation of Ribosomal Proteins. Curr Biol [Internet]. 2007 May 1 [cited 2020 Jul 28];17(9):749–60. Available from: https://pubmed.ncbi.nlm.nih.gov/17446074/
38. Mathieson T, Franken H, Kosinski J, Kurzawa N, Zinn N, Sweetman G, et al. Systematic analysis of protein turnover in primary cells. Nat Commun [Internet]. 2018 Dec 1 [cited 2020 Jul 28];9(1):689–99. Available from: /pmc/articles/PMC5814408/?report=abstract
39. Meng L, Tong J, Wang H, Tao C, Wang Q, Niu C, et al. PPE38 Protein of Mycobacterium tuberculosis Inhibits Macrophage MHC Class I Expression and Dampens CD8+ T Cell Responses. Front Cell Infect Microbiol [Internet]. 2017 [cited 2019 Apr 15];7:68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28348981
40. Mannervik B. Molecular enzymology of the glyoxalase system [Internet]. Vol. 23, Drug Metabolism and Drug Interactions. Freund Publishing House Ltd; 2008 [cited 2021 Mar 3]. p. 13–27. Available from: https://pubmed.ncbi.nlm.nih.gov/18533362/
41. Thornalley PJ. Protecting the genome: Defence against nucleotide glycation and emerging role of glyoxalase I overexpression in multidrug resistance in cancer chemotherapy. In: Biochemical Society Transactions [Internet]. Portland Press Ltd; 2003 [cited 2021 Mar 3]. p. 1372–7. Available from: https://pubmed.ncbi.nlm.nih.gov/14641066/
42. Thornalley PJ. The glyoxalase system: New developments towards functional characterization of a metabolic pathway fundamental to biological life [Internet]. Vol. 269, Biochemical Journal. Biochem J; 1990 [cited 2021 Mar 3]. p. 1–11. Available from: https://pubmed.ncbi.nlm.nih.gov/2198020/
43. Thornalley PJ, Jahan I, Ng R. Suppression of the Accumulation of Triosephosphates and Increased Formation of Methylglyoxal in Human Red Blood Cells during Hyperglycaemia by Thiamine In Vitro. J Biochem [Internet]. 2001 Apr 1 [cited 2021 Mar 3];129(4):543–9. Available from: https://academic.oup.com/jb/article-lookup/doi/10.1093/oxfordjournals.jbchem.a002889
44. Jin X, Yao T, Zhou Z, Zhu J, Zhang S, Hu W, et al. Advanced glycation end products enhance macrophages polarization into M1 phenotype through activating RAGE/NF-B Pathway. Biomed Res Int. 2015;2015:1–12.
45. Torres-Huerta A, Villaseñor T, Flores-Alcantar A, Parada C, Alemán-Navarro E, Espitia C, et al. Interaction of the CD43 Sialomucin with the Mycobacterium tuberculosis Cpn60.2 Chaperonin Leads to Tumor Necrosis Factor Alpha Production. Infect Immun [Internet]. 2017 [cited 2019 Sep 1];85(3):e00915-16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28069816
46. Ramírez-Pliego O, Escobar-Zárate DL, Rivera-Martínez GM, Cervantes-Badillo MG, Esquivel-Guadarrama FR, Rosas-Salgado G, et al. CD43 signals induce Type One lineage commitment of human CD4+ T cells. BMC Immunol [Internet]. 2007 Nov 23 [cited 2019 Sep 1];8(1):30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18036228
47. Galindo-Albarrán AO, Ramírez-Pliego O, Labastida-Conde RG, Melchy-Pérez EI, Liquitaya-Montiel A, Esquivel-Guadarrama FR, et al. CD43 signals prepare human T cells to receive cytokine differentiation signals. J Cell Physiol [Internet]. 2014 Feb [cited 2019 Sep 1];229(2):172–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24328034
48. Zhang Q, Lenardo MJ, Baltimore D. 30 Years of NF-kB: A Blossoming of Relevance to Human Pathobiology. Cell [Internet]. 2017 [cited 2019 Sep 1];168:37–57. Available from: http://dx.doi.org/10.1016/j.cell.2016.12.012

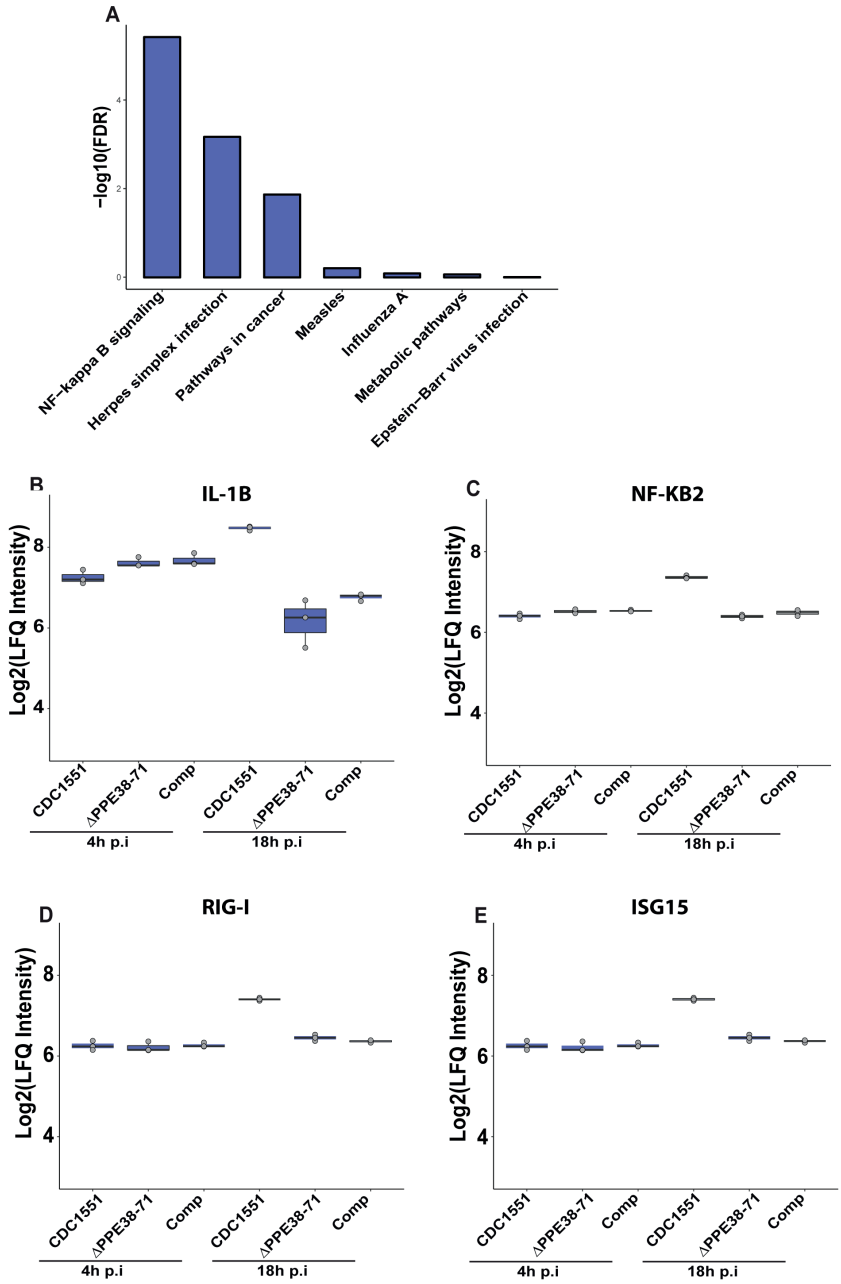
49. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther [Internet]. 2017 Jul 14 [cited 2019 Mar 29];2:17023. Available from: http://www.nature.com/articles/sigtrans201723
50. Manzanillo PS, Shiloh MU, Portnoy DA, Cox JS. Mycobacterium tuberculosis activates the DNA-dependent cytosolic surveillance pathway within macrophages. Cell Host Microbe [Internet]. 2012 May 17 [cited 2019 Mar 25];11(5):469–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22607800
51. Lernbecher T, Müller U, Wirth T. Distinct NF-κB/Rel transcription factors are responsible for tissue-specific and inducible gene activation. Nature. 1993;365(6448):767–70.
52. Do RKG, Hatada E, Lee H, Tourigny MR, Hilbert D, Chen-Kiang S. Attenuation of apoptosis underlies B lymphocyte stimulator enhancement of humoral immune response. J Exp Med. 2000 Oct 2;192(7):953–64.
53. Jiang HY, Petrovas C, Sonenshein GE. RelB-p50 NF-B Complexes Are Selectively Induced by Cytomegalovirus Immediate-Early Protein 1: Differential Regulation of Bcl-x L Promoter Activity by NF-B Family Members. J Virol. 2002;76(11):5737–47.
54. Sun S-C. The non-canonical NF-κB pathway in immunity and inflammation. Nat Rev Immunol [Internet]. 2017 Jun 5 [cited 2019 Mar 26];17(9):545–58. Available from: http://www.nature.com/doifinder/10.1038/nri.2017.52
55. Weih F, Durham SK, Barton DS, Sha WC, Baltimore D, Bravo R. p50-NF-κB complexes partially compensate for the absence of RelB: Severely increased pathology in p50(-/-)relB(-/-) double-knockout mice. J Exp Med. 1997 Apr 7;185(7):1359–70.
56. Weih F, Carrasco D, Durham SK, Barton DS, Rizzo CA, Ryseck RP, et al. Multiorgan inflammation and hematopoietic abnormalities in mice with a targeted disruption of RelB, a member of the NF-κB/Rel family. Cell. 1995 Jan 27;80(2):331–40.
57. Gasparini C, Foxwell B, Feldmann M. RelB/p50 regulates TNF production in LPS-stimulated dendritic cells and macrophages. Cytokine [Internet]. 2013 Mar 1 [cited 2019 Apr 18];61(3):736–40. Available from: https://www-sciencedirect-com.vu-nl.idm.oclc.org/science/article/pii/S1043466613000276
58. Cash E, Minty A, Ferrara P, Caput D, Fradelizi D, Rott O. Macrophage-inactivating IL-13 suppresses experimental autoimmune encephalomyelitis in rats. J Immunol [Internet]. 1994 Nov 1 [cited 2019 Dec 17];153(9):4258–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7523520
59. Cappelli G, Volpe E, Grassi M, Liseo B, Colizzi V, Mariani F. Profiling of Mycobacterium tuberculosis gene expression during human macrophage infection: Upregulation of the alternative sigma factor G, a group of transcriptional regulators, and proteins with unknown function. Res Microbiol. 2006 Jun;157(5):445–55.
60. Fontán P, Aris V, Ghanny S, Soteropoulos P, Smith I. Global transcriptional profile of Mycobacterium tuberculosis during THP-1 human macrophage infection. Infect Immun. 2008 Feb;76(2):717–25.
61. Li AH, Waddell SJ, Hinds J, Malloff CA, Bains M, Hancock RE, et al. Contrasting transcriptional responses of a virulent and an attenuated strain of Mycobacterium tuberculosis infecting macrophages. PLoS One. 2010;5(6):e11066.
62. Sayes F, Sun L, Di Luca M, Simeone R, Degaiffier N, Fiette L, et al. Strong immunogenicity and cross-reactivity of Mycobacterium tuberculosis ESX-5 type VII secretion-encoded PE-PPE proteins predicts vaccine potential. Cell Host Microbe. 2012 Apr 19;11(4):352–63.

63. Ates LS, van der Woude AD, Bestebroer J, van Stempvoort G, Musters RJP, Garcia-Vallejo JJ, et al. The ESX-5 System of Pathogenic Mycobacteria Is Involved In Capsule Integrity and Virulence through Its Substrate PPE10. Behr MA, editor. PLOS Pathog [Internet]. 2016 Jun 9 [cited 2019 Dec 19];12(6):e1005696. Available from: http://dx.plos.org/10.1371/journal.ppat.1005696
64. Weih F, Warr G, Yang H, Bravo R. Multifocal defects in immune responses in RelB-deficient mice. J Immunol [Internet]. 1997 Jun 1 [cited 2019 Apr 11];158(11):5211–8. Available from: http://www.jimmunol.org/content/158/11/5211.short
65. McEvoy CRE, van Helden PD, Warren RM, van Pittius N, Gey van Pittius NC. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. BMC Evol Biol [Internet]. 2009 Jan;9(1):237. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-237
66. Madacki J, Orgeur M, Fiol GM, Frigui W, Ma L, Brosch R, et al. ESX-1-Independent Horizontal Gene Transfer by Mycobacterium tuberculosis Complex Strains Downloaded from. 2021 [cited 2021 May 24];0(0):e00965-21. Available from: http://mbio.asm.org/
67. Mouton JM, Heunis T, Dippenaar A, Gallant JL, Kleynhans L, Sampson SL. Comprehensive Characterization of the Attenuated Double Auxotroph Mycobacterium tuberculosisΔleuDΔpanCD as an Alternative to H37Rv. Front Microbiol [Internet]. 2019 Aug 20 [cited 2020 Nov 24];10(AUG):1922. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2019.01922/full
68. Gallant J, Mouton J, Ummels R, ten Hagen-Jongman C, Kriel N, Pain A, et al. Identification of gene fusion events in Mycobacterium tuberculosis that encode chimeric proteins. NAR Genomics Bioinformatics. 2020 Jun 1;2(2).
69. Stapels DAC, Hill PWS, Westermann AJ, Fisher RA, Thurston TL, Saliba AE, et al. Salmonella persisters undermine host immune defenses during antibiotic treatment. Science (80- ). 2018 Dec 7;362(6419):1156–60.
70. Manca C, Tsenova L, Bergtold A, Freeman S, Tovey M, Musser JM, et al. Virulence of a Mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-α/β. Proc Natl Acad Sci [Internet]. 2001 May 8;98(10):5752 LP – 5757. Available from: http://www.pnas.org/content/98/10/5752.abstract
71. Loeb KR, Haas AL. The interferon-inducible 15-kDa ubiquitin homolog conjugates to intracellular proteins. J Biol Chem. 1992 Apr;267(11):7806–13.
72. Reed MB, Domenech P, Manca C, Su H, Barczak AK, Kreiswirth BN, et al. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature [Internet]. 2004 Sep 2 [cited 2019 Apr 18];431(7004):84–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15343336
73. Sinsimer D, Huet G, Manca C, Tsenova L, Koo M-S, Kurepina N, et al. The phenolic glycolipid of Mycobacterium tuberculosis differentially modulates the early host cytokine response but does not in itself confer hypervirulence. Infect Immun. 2008 Jul;76(7):3027–36.
74. Sousa J, Cá B, Maceiras AR, Simões-Costa L, Fonseca KL, Fernandes AI, et al. Mycobacterium tuberculosis associated with severe tuberculosis evades cytosolic surveillance systems and modulates IL-1β production. Nat Commun. 2020 Dec 1;11(1):1–14.

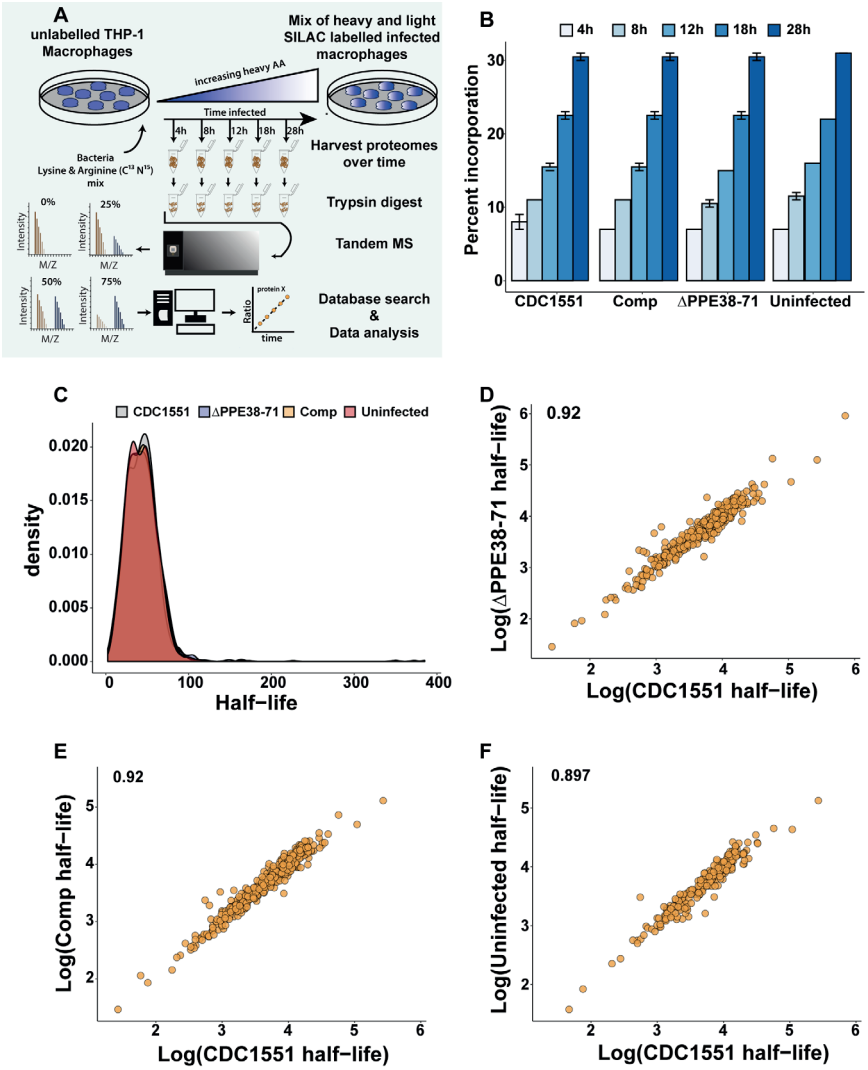
# SUPPLEMENTARY FIGURES



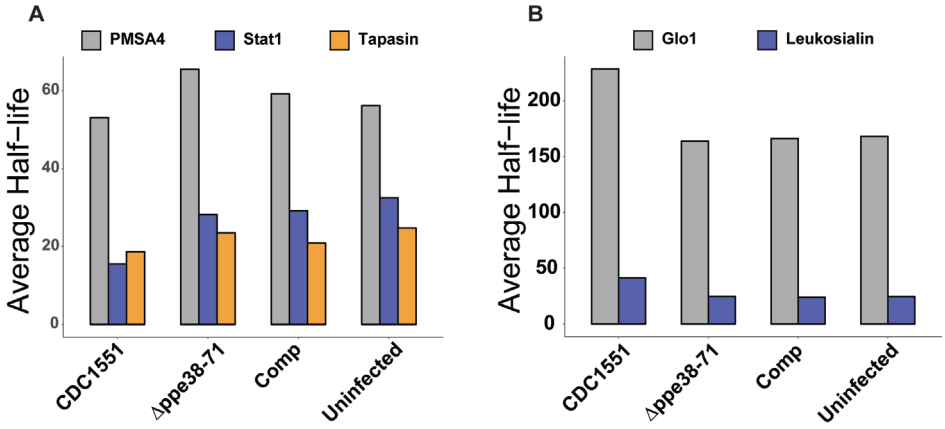

**Supplementary figure 1: Quality control of label -free mass spectrometry data, related to figure 2: A)** 2 Boxplot representing LFQ intensities across all samples. Significant deviations from the mean 3 between samples were ruled out by hypothesis testing using one-way ANOVA and Tukey HSD 4 post-hoc testing, q-value was set at 0.05. **B)** Principle component analysis was used to assess 5 the clustering of groups and replicates. Little to no clear separation occurs at 4h post-infection 6 between all groups. **C)** Separation was observed between all three groups at 18h post-infection 7 and clustered based on strain genotype. Separation occurs between M. tuberculosis CDC1551- 8 infected macrophages at 4h and 18h post -infection on the first component while replicates 9 separated on the second component in all three strains namely, **D)** M. tuberculosis CDC1551, 10 **E)** Δppe38-71 and **F)** complement-infected macrophages. Data used for all samples is 11 represented by 3 biological replicates per sample.



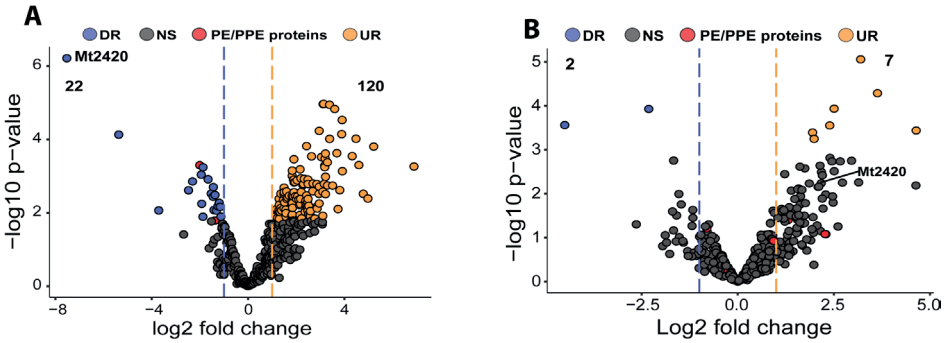

**Supplementary figure 2: Enrichment analysis indicates altered inflammatory responses in M. tuberculosis-infected macrophages, related to figure 2. A)** Gene ontology enrichment using the WebGestalt server. Significantly regulated proteins and their respective effect sizes, as determined in Figure 3, were used as the gene set input and enriched against the Kyoto Encyclopaedia of Genes and Genomes and scaled using the false-discovery rate (FDR). Log2 LFQ intensities for **B)** IL-1B, **C)** NFkB2, **D)** RIG-I and **E)** ISG15 identified from macrophages infected with the different strains at the indicated time points. Individual points in each group represent each biological replicate, while error bars represent the standard deviation.

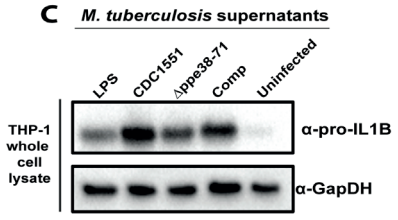


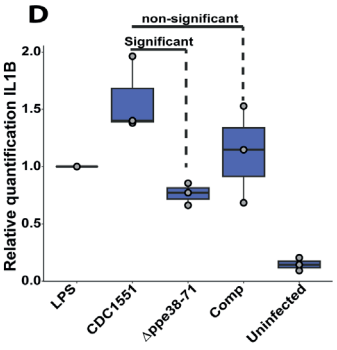

**Supplementary figure 3. Global protein turnover in THP-1 macrophages is not significantly altered during infection, related to figure 3. A)** A pulse SILAC labelling approach was used to obtain protein half-lives. Proteins were harvested at 4h, 8h, 12h, 18h and 28h post-infection from uninfected THP-1 macrophages as well as macrophages infected with M. tuberculosis CDC1551, M. tuberculosis Δppe38-71 and the complemented strain. **B)** Incorporation of "heavy" arginine and lysine over time. Maximal incorporation of ~30% was achieved in all infected macrophages and in the uninfected control. **C)** Density plot representing the distribution of protein half-lives from both infected and uninfected macrophages. Half-lives were calculated from the mean raw H/L ratios of two independent experiments. **D-F).** Multiple scatterplots of pairwise comparisons between conditions. The points of the scatter plot represent from log2 transformed half-lives, with Pearson correlation coefficients for each comparison.




**Supplementary figure 4: Proteins involved in response to infection that have differentially regulated half-lives between THP-1 macrophages infected with M. tuberculosis CDC1551 and M. tuberculosis Δppe38-71, related to figure 4:** From the differentially regulated half-lives, selected proteins with a **A)** rapid or **B)** slow turnover in THP-1 macrophages infected with M. tuberculosis CDC 1551 compared to other conditions are displayed. Values are representative of two independent pSILAC experiments.



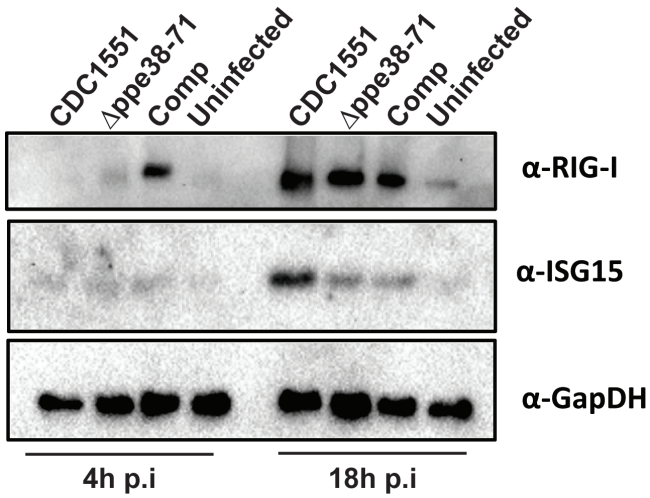



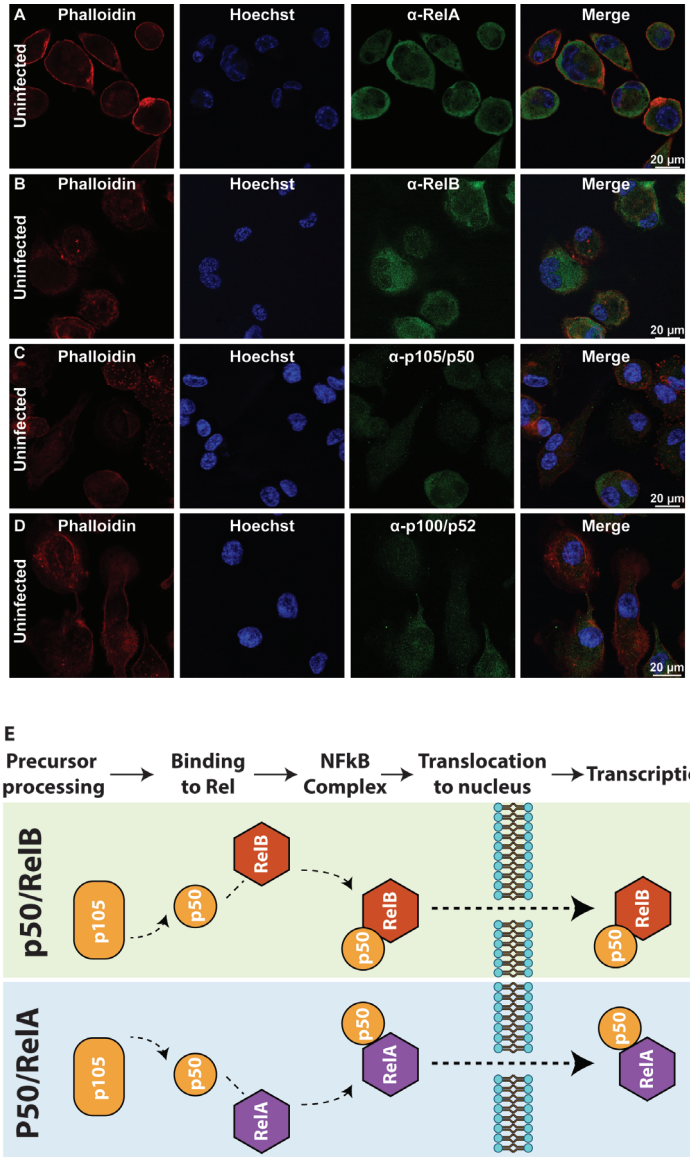

**E** One way Anova + Tukey HSD posthoc

| Strain | diff | lwr | upr | p.adj |
|---|---|---|---|---|
| Comp-CDC1551 | -0.46 | -1.12 | 0.20 | 0.22 |
| Δppe38-71-CDC1551 | -0.82 | -1.48 | -0.16 | 0.01 |
| LPS-CDC1551 | -0.58 | -1.24 | 0.08 | 0.09 |
| Uninfected-CDC1551 | -1.43 | -2.09 | -0.78 | 0.00 |
| dPPE38-Comp | -0.36 | -1.01 | 0.30 | 0.43 |
| LPS-Comp | -0.12 | -0.78 | 0.54 | 0.97 |
| Uninfected-Comp | -0.97 | -1.63 | -0.31 | 0.00 |
| LPS-Δppe38-71 | 0.24 | -0.42 | 0.90 | 0.76 |
| Uninfected-Δppe38-71 | -0.62 | -1.27 | 0.04 | 0.07 |
| Uninfected-LPS | -0.85 | -1.51 | -0.19 | 0.01 |

**Supplementary figure 5: Label-free mass spectrometry analysis of detergent free M. tuberculosis supernatants shows restored complementation but no PE/PPE proteins, related to figure 5.** Volcano plots depicting differential expression of proteins found in the detergent-free supernatants from **A)** M. tuberculosis Δppe38-71 and **B**) complemented strain compared to M. tuberculosis CDC1551. A q-value cut-off was set to 0.05 and the log fold change was set to 1. Data is representative of three independent experiments. THP-1 macrophages were stimulated with cell-free supernatants from M. tuberculosis CDC1551, Δppe38-71 or complemented strains. Stimulation with lipopolysaccharides from E. coli served as a positive control and unstimulated macrophages as negative control. Macrophages were lysed and probed for IL-1B expression using **C)** Western blot, which was quantified by **D)** densitometry. The data is a representative of three independent experiments and statistical significance was determined using **E)** one-way ANOVA, followed by a Tukey HSD post-hoc test with a q-value set at 0.05




**Supplementary figure 6: Expression of ISG15 and RIG-I at 18 hours post infection shows no differential regulation in RIG-I but upregulation of ISG15 in M. tuberculosis CDC1551, related to figure 6.** THP-1 macrophages were infected with M. tuberculosis CDC1551, M. tuberculosis Δppe38-71 and the complemented strain and probed for RIG-I and ISG15 at 4 hours and 18 hours post infection by Western blot. GapDH was used as a loading control and uninfected macrophages were used as a stimulus control.




**Supplementary figure 7: Confocal microscopy of RelA and RelB in control samples, related to figure 7.** Representative images of uninfected THP-1 macrophages used as control in for the experiment depicted in Figure 7. Uninfected macrophages were labelled with phalloidin (F-actin), the nucleus was stained with Hoechst and **A)** RelA, **B)** RelB, **C)** p105/p50 and **D)** p100/p52 conjugated-Alexafluor 488 was used to detect NFkB proteins. **E)** Schematic illustration of the NF-KB pathway stimulated by infection with M. tuberculosis CDC1551 and M. tuberculosis Δppe38-71.

# 6

# Investigating non-sterilizing cure in TB patients at the end of successful anti-TB therapy.

Caroline G.G. Beltran[1,2*]
Tiaan Heunis[1,2,3]
James Gallant[1,2,4]
Rouxjeane Venter[1,2]
Andre G. Loxton[1,2]
Matthias Trost[3]
Nelita du Plessis[1,2]
Jill Winter[4]
Stephanus T. Malherbe[1,2]
Bavesh D. Kana[1,6,7]
Gerhard Walzl[1,2]

[1]Department of Science and Technology/National Research Foundation, Centre of Excellence for Biomedical Tuberculosis Research and South African Medical Research Council Centre for Tuberculosis Research, Cape Town, South Africa
[2]Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Department of Biomedical Sciences, Stellenbosch University, Cape Town, South Africa
[3]Faculty of Medical Sciences, Biosciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom
[4]Section Molecular Microbiology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[5]Catalysis Foundation for Health, San Ramon, CA, USA
[6] DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Health Sciences, School of Pathology, University of the Witwatersrand and the National Health Laboratory Service, Johannesburg, South Africa
[7]MRC-CAPRISA HIV-TB Pathogenesis and Treatment Research Unit, Centre for the AIDS Programme of Research in South Africa, CAPRISA, Durban, South Africa

Frontiers in cellular and infection microbiology

DOI: https://doi.org/10.3389/fcimb.2020.00443

## ABSTRACT

*Mycobacterium tuberculosis* (Mtb) is extremely recalcitrant to antimicrobial chemotherapy requiring 6 months to treat drug-sensitive tuberculosis (TB). Despite this, 4–10% of cured patients will develop recurrent disease within 12 months after completing therapy. Reasons for relapse in cured TB patients remains speculative, attributed to both pathogen and host factors. Populations of dormant bacilli are hypothesized to cause relapse in initially cured TB patients however, development of tests to convincingly demonstrate their presence at the end of anti-TB treatment has been challenging. Previous studies have indicated the utility of culture filtrate supplemented media (CFSM) to detect differentially culturable tubercle bacilli (DCTB). Here, we show that 3/22 of clinically cured patients retained DCTB in induced sputum and bronchoalveolar lavage fluid (BALF), with one DCTB positive patient relapsing within the first year of completing therapy. We also show a correlation of DCTB status with "unresolved" end of treatment FDG PET-CT imaging. Additionally, 19 end of treatment induced sputum samples from patients not undergoing bronchoscopy were assessed for DCTB, identifying a further relapse case with DCTB. We further show that induced sputum is a less reliable source for the DCTB assay at the end of treatment, limiting the utility of this assay in a clinical setting. We next investigated the host proteome at the site of disease (BALF) using multiplexed proteomic analysis and compared these to active TB cases to identify host-specific factors indicative of cure. Distinct signatures stratified active from cured TB patients into distinct groups, with a DCTB positive, subsequently relapsing, end of treatment patient showing a proteomic signature closer to active TB disease than cure. This exploratory study offers evidence of live Mtb, undetectable with conventional culture methods, at the end of clinically successful treatment and putative host protein biomarkers of active disease and cure. These findings have implications for the assessment of true sterilizing cure in TB patients and opens new avenues for targeted approaches to monitor treatment response.

## INTRODUCTION

*Mycobacterium tuberculosis* (Mtb) is a highly complex and well-adapted pathogen that causes tuberculosis (TB). The success of this pathogen can be attributed to its ability to evade the protective host immune response and its recalcitrance to antimicrobial chemotherapy. Although cure of the majority of patients treated with the standard 6 month multidrug regimen indicates that treatment is highly effective, ~4–10% of clinically cured patients will develop recurrent disease within the first 12 months after completing therapy (1,2). Recurrence can fall into two categories; relapse due to the reactivation of the original infecting strain or reinfection due to exogenous infection with a strain of Mtb different from the original infecting strain (2). In high endemic countries, the line between reinfection and relapse is often blurred since patients are likely to be continually exposed to similar strains (3).

Current observed relapse rates indicate that a fraction of patients, whilst appearing clinically cured, still harbour a sub-clinical Mtb infection that failed to be sterilized by current treatment regimens. Indeed, there is growing evidence pointing to residual live Mtb in patients at the end of treatment (4,5). Previous work from our group has shown that most patients at the end of treatment retain 18F-fluorodyoxyglucose positron emission tomography (FDG-PET) avid lung lesions indicative of persisting inflammation, even after sputum cultures are consistently negative. Furthermore, Mtb mRNA can be detected in sputum and bronchoalveolar lavage fluid (BALF) in a significant fraction of cured patients at the end of curative treatment, possibly indicating residual live Mtb (4). It is well-established that Mtb, amongst other bacteria, can form physiologically heterogeneous populations both *in-vitro* and *in-vivo* (6–11). In patients, a sub-population of metabolically distinct bacilli, defined as differentially culturable tubercle bacilli (DCTB) can be detected in sputa when grown in the presence of sterilized culture filtrate supplemented media (CFSM). This sub-population appears to become more prominent during early chemotherapy (12,13), yet is not directly detectable by conventional microbiological diagnostic methods. It remains to be conclusively determined whether the presence of DCTB cells correlates with clinical treatment response and relapse rates. Host-specific factors are likely to also be critical in determining favourable or unfavourable outcomes even in the presence of persistent bacteria.

Part of the complexity of the problem lies with the host itself, where, paradoxically, anatomic niches in the form of granulomas sequester infectious bacteria from the lung and favour the development and long-term survival of dormant Mtb, subsequently giving rise to actively replicating bacterial populations and dissemination (14–17). The

link between dormancy and reactivation of disease has been demonstrated in animal models for several bacterial species, including Mtb (18–21).

The aim of this study was to investigate whether patients at the end of successful anti-TB treatment retain live, DCTB and to correlate this with FDG PET-CT activity. We used CFSM to culture mycobacteria from both BALF and induced sputum. We also analysed BALF supernatant from these same clinically cured patients and an equal number of active TB samples, using shotgun proteomics. The results provide a global picture of the host proteome at the site of disease and gives insight into host responses to active TB infection vs. clinical cure. The finding that DCTB and host biomarkers for active TB may be present in cured TB patients, albeit in a small proportion of patients, opens new avenues for targeted approaches to monitor treatment response.

## MATERIALS AND METHODS

A summary workflow diagram of the sample recruitment and methods is presented in Figure 1. Twenty-two patients were included for a full characterization at the end of treatment (EOT). Patients were scanned by FDG PET-CT imaging and had a sputum induction by inhalation of hypertonic saline. These patients also underwent a bronchoscopy where a bronchoalveolar lavage fluid (BALF) and post-bronchoscopy sputum (PB sputum) was collected. The induced sputum, PB sputum and BALF pellet were assessed for DCTB at EOT and to correlate this with FDG PET-CT lesion activity. A subset of the BALF supernatant samples (n = 5) were used to develop a protocol that utilized proteomics analysis to provide insight into the global host proteomics changes at the site of disease. A further nineteen EOT patients were assessed for DCTB in their induced sputum, only.

### Patient Recruitment

Ethical consent was obtained from the Medical Human Research Ethics committee (HREC) of Stellenbosch University (HREC references: N10/01/013 and N16/05/070). All methods were performed in accordance with the relevant guidelines and regulations. EOT patients were included in the study if they met the following criteria: willing to give consent, willing to have their HIV status tested, aged between 18 and 70 years, not taking any corticosteroid medication, having undergone successful anti-TB treatment (confirmed by two consecutive, negative sputum cultures). Detailed clinical medical history was taken at baseline (history of smoking, previous TB, other medication) and physical tests included chest X-rays and blood test for HbA1c level. All patients were given standard anti-TB treatment consisting of 2 months intensive phase [rifampicin

(RIF), isoniazid (INH), ethambutol and pyrazinamide] and 4 months of RIF and INH as continuation phase. After informed consent, 41 EOT patients were recruited from studies that took place in primary health care clinics in Northern Cape Town. Patients were followed up to 1 year after the end of treatment (M18) to assess recurrence. Ten healthy house-hold contacts (negative controls) from the same community and two pre-treatment active TB cases (positive controls) were included as control subjects for the DCTB assay. A further five pre-treatment active TB cases were recruited for proteomics analysis.

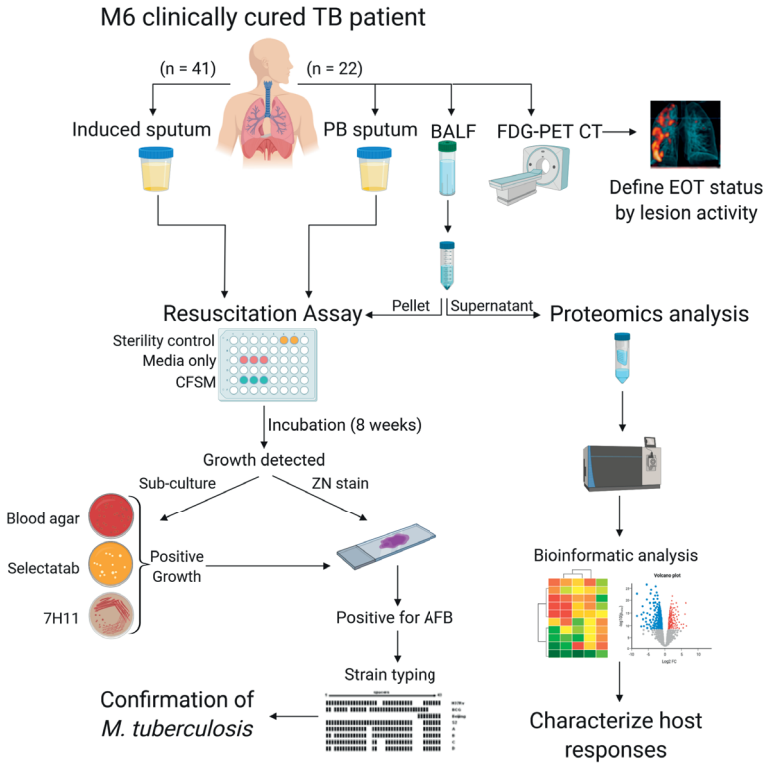



**Figure 1:** Workflow diagram of the sample processing and analysis to investigate sterilizing cure in tuberculosis (TB) patients at the end of successful anti-TB therapy. Three sample types were analysed for resuscitation of M. tuberculosis; an induced sputum (n=41), bronchoscopy with bronchoalveolar lavage fluid (BALF)(n=22) and post-bronchoscopy (PB) sputum (n=22) were collected. These patients were also scanned using FDG-PET CT to classify lesion activity at the end of treatment. The BALF supernatant from five patients was used for multiplexed proteomics analysis using tandem mass tag (TMT 10 plex) and compared to the proteome of active TB cases. The resuscitation assay consisted of a series of confirmatory tests once growth was detected including sub-culturing on blood agar, Mycobacterium Selectatab and Middlebrook 7H11 Agar. Colonies corresponding to typical growth of M. tuberculosis were subsequently picked and confirmed for acid fast bacilli (AFB) using ZN staining before being strain typed using spoligotyping.

## End of Treatment Induced Sputum

Sputum was induced with 20 mL of 3% hypertonic saline solution delivered through ultrasonic nebulizer and collected in a sterile sputum container. Sputum samples were stored at 4°C prior to processing (sputa were processed within 24 h of collection to minimize growth of commensal bacteria).

## FDG PET-CT

Patients were scanned at the Western Cape Academic Positron Emission Tomography (PET) Centre, at Tygerberg Academic Hospital. All patients received between 5 and 7 mCi of intravenous 18F FDG (18-F Fluorodeoxyglucose), which is taken up by metabolically active cells and indicates inflammation. Patients were scanned 1 h after the injection using a Philips Gemini scanner. Whole lung analysis was performed on the scans to quantify the total glycolytic activity index (TGAI) on PET (indicating total inflammatory burden) using lesion-free lung as a background (4), the most intense residual lesion uptake on PET, (3) the total residual cavity volume on CT.

## EOT Bronchoscopy and BALF Collection

One week following the FDG PET-CT scan and induced sputum visit, an experienced bronchoscopist conducted a bronchoscopy and bronchoalveolar lavage on patients at the Pulmonology Unit, Tygerberg Academic Hospital. Up to 200 mL of sterile physiologic saline solution were instilled (in divided aliquots of 20–60 mL each) and subsequently aspirated to obtain BALF from the lobe that showed the most residual infiltrates on an accompanying chest X-ray. BALF and the induced sputum samples were split into two volumes following collection and immediately processed. A post-bronchoscopy (PB) sputum was collected immediately after the bronchoscopy for the DCTB assay.

## Sputum and BALF Decontamination

BALF samples (50 mL) were concentrated by centrifugation at 4000 × g for 20 min at 10°C, the supernatant transferred to a fresh tube and stored at −80°C for proteomics analysis, leaving ~5 mL pellet. The BALF pellets and sputa were decontaminated by adding equal volumes of N-acetyl-L-cysteine- 2% sodium hydroxide (NALC-NaOH) using the commercial MycoPrep™ kit (BD Biosciences, USA) and incubating for 15 min at room temperature (RT) with brief vortexing every 5 min. Specimens were neutralized with phosphate buffer saline (PBS) (33 mM Na2HPO4, 33 mM KH2PO4; pH 6.8) and concentrated by centrifugation at 4,000 × g for 20 min at 10°C. The pellets were resuspended in 300 µL PBS and used immediately for the DCTB assay.

## Preparation of Culture Filtrate Supplemented Media

The laboratory strain of Mtb, H37Rv, was used to produce culture filtrate (CF), as described previously (22). Briefly, Mtb cultures were grown in Middlebrook 7H9 medium containing 0.05% tween and supplemented with 10% OADC (oleic acid, albumin, dextrose, and catalase; BD Biosciences) in vented flasks and kept stationary at 37°C until mid-exponential phase was reached (OD600 nm 0.6–1.0). Cultures were centrifuged at 4,000 × g for 15 min and supernatants were sterilized by double filtration using sterile, disposable Durapore Membrane Filters PVDF (0.22 µm pore size). CF was supplemented with fresh 7H9 media (at a 1:1 ratio) and polymyxin B, amphotericin B, nalidixic acid, trimethoprim, and azlocillin (PANTA™ antibiotic mixture, BD Biosciences) were added to increase subsequent selectivity for Mtb growth. A 450 µL aliquot of the CFSM was added to a 48-well cell culture plate (Nunc, Thermo Scientific). Sterility of the CF was verified by including neat aliquots in control wells of the cell culture plate.

## Addition of Sample to 48-Well Culture Plate

Decontaminated induced sputum and BALF samples (50 µL) were added in triplicate to each well containing CFSM. An additional three wells containing 7H9 media (without CFSM) were also included as a media control comparison and the same volume of samples was added to each of these wells (see DCTB assay for plate layout in Figure 1). For the blank controls, 50 µL of PBS was added to the CFSM. Culture plates were sealed with micropore tape and incubated without shaking at 37°C for a period of 8 weeks. Plates were checked for growth weekly.



## Confirmation of M. tuberculosis Growth

Confirmatory assays were conducted once growth was identified visually (through observation of turbidity), to confirm the presence of Mtb, as opposed to contamination by other microorganisms. Confirmatory testing included a combination of Ziehl-Neelsen (ZN) staining and sub-culturing onto; blood agar plates to check for contamination; Mycobacterium Selectatab media (Mast Group) to isolate single colonies and Middlebrook 7H11 Agar (BD Life Sciences). Once colonies corresponding to typical growth of Mtb (non-pigmented, rough, dry colonies) were detected on Selectatab media and/or 7H11 media, a colony was picked for ZN staining as well as resuspended in 500 µL ultrapure water (MilliQ, Merck) and heat inactivated at 100°C for 30 min to extract DNA for strain typing using spoligotyping. Replicate wells for both CFSM and media control were followed up independently. A negative culture well was defined as one in which neither Mtb grew, nor contamination was observed. Spoligotyping was conducted as previously described (23), and the recovered strain was matched to the original infecting strain isolated from baseline sputum cultures.

## Proteomics Analysis

### Samples

For the proteomics analysis of BALF supernatant, five EOT patients also analysed for DCTB (Figure 1), were chosen (including PID338 who subsequently relapsed) and BALF from five active TB (before treatment initiation) patients were included as a comparison. The active TB patients were recruited at their baseline visit and underwent a bronchoscopy, as described above. After cell collection, the BALF supernatants (15 mL) were filter sterilized using 0.22 µm PVDF filters prior to processing. The filtered BALF samples were concentrated using Amicon ultra-15 kDa centrifugal filters (Merck) to ~2.5 mL and depleted using the ProteoPrep Blue Albumin and IgG Depletion kit (Sigma-Aldrich), as recommended by the supplier. Ice-cold acetone was added to the depleted samples and proteins were precipitated overnight at −20°C. Samples were centrifuged at 14 000 × g for 10 min and the supernatant removed. The protein pellets were resuspended in 50 µL 8 M urea in 100 mM triethylammonium bicarbonate (TEAB) and sonicated on ice for 10 min to completely reconstitute the protein. Protein concentration was determined using a Bradford assay.

### Proteolytic Digestion and TMT Labelling of Peptides

Aliquots corresponding to 50 µg protein from each sample were used for filter-aided sample preparation (FASP) (24). Samples were reduced with 5 mM tris(2-carboxyethyl) phosphine (TCEP) for 1 h at RT and alkylated with 5.5 mM iodoacetamide (IAA) for 1 h in the dark at RT. Samples were concentrated using Amicon 30 kDa Ultra 0.5 mL spin filters (Merck) at 14 000 × g for 15 min before adding 400 µL 50 mM TEAB to each spin filter, and further centrifugation at 14 000 × g for 15 min. This process was repeated a further three times, discarding the flow through between washes. Following the wash steps, 400 µL of 50 mM TEAB was added directly onto each filter and the samples centrifuged at 14 000 × g for 15 min. This process was repeated a further two times. Modified sequencing grade trypsin (Promega) was added at a 1:50 ratio of trypsin to protein to each filter and samples were incubated at 37°C for 18 h. Following trypsin digestion, filters were transferred to a new collection tube and centrifuged at 14 000 × g for 15 min, collecting peptides in the flow through. Peptides were further eluted using 400 µL of 100 mM sodium chloride (NaCl) and centrifuged again. Peptide eluates were pooled and concentrated by vacuum drying. Peptides were subsequently desalted using in-house packed C18 STAGE tips, vacuum dried, and stored until labelling with tandem mass tags. TMT 10 plex (Thermo Fisher Scientific) labelling was done according to manufacturer's instructions. Briefly, peptide samples were reconstituted in 100 µL 100 mM TEAB and the TMT labels were reconstituted in 41 µL anhydrous acetonitrile. The labelling reaction was quenched using 8 µL of 5% hydroxylamine and incubating for 15

min at RT. Approximately 15 µg from each sample was combined into a new tube and desalted and fractionated using a C18 STAGE tip. STAGE tip-based fractionation was done using increasing gradients of acetonitrile in 10 mM TEAB as follows; 7.5, 10, 12.5, 15.5, 17.5, 20, 22.5, 25, 30, 50% and collecting the eluates between each gradient into individual tubes before vacuum drying.

### Mass Spectrometry

Peptides were dissolved in 2% acetonitrile containing 0.1% trifluoroacetic acid, and each sample was independently analysed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific), connected to an UltiMate 3000 RSLCnano System (Thermo Fisher Scientific). Peptides were injected on an Acclaim PepMap 100 C18 LC trap column (100 µm ID ×20 mm, 3 µm, 100 Å) followed by separation on an EASY-Spray nanoLC C18 column (75 µm ID ×500 mm, 2 µm, 100 Å) at a flow rate of 300 nl.min−1. Solvent A was water containing 0.1% formic acid, and solvent B was 80% acetonitrile containing 0.1% formic acid. The gradient used for analysis was as follows: solvent B was maintained at 3% for 5 min, followed by an increase from 3 to 35% B in 90 min, 35–90% B in 0.5 min, maintained at 90% B for 4 min, followed by a decrease to 3% in 0.5 min and equilibration at 3% for 10 min. The Orbitrap Fusion Lumos was operated in positive-ion data-dependent mode. Precursor ion (MS1) scans were performed in the Orbitrap mass analyser in the range of 375–1,500 m/z, with a resolution of 120 000 at 200 m/z. An automatic gain control (AGC) target of $4 \times 10^5$ and an ion injection time of 50 ms was allowed. Precursor ions were isolated using a quadrupole mass filter with an isolation width of 0.7 m/z, and fragmented using collision-induced dissociation (CID) with a collision energy of 35%. MS2 spectra were acquired in the linear ion (IT) trap using turbo mode. An AGC target of $1 \times 10^4$ and a maximum injection time of 50 ms was allowed. Ten MS2 fragment ions were co-selected for MS3 analysis using synchronous precursor selection (SPS), and underwent high-energy collisional dissociation (HCD) with collision energy set to 65% to ensure maximal TMT reporter ion yield. SPS-MS3 fragment ions were analysed in the Orbitrap mass analyser at a resolution of 60 000 at 200 m/z. An AGC target of $5 \times 10^4$ and maximum injection time of 120 ms was allowed. The number of MS2 and MS3 events between full scans was determined on-the-fly to maintain a 3 s fixed duty cycle. Dynamic exclusion of ions within a ±10 ppm m/z window was implemented using a 35 s exclusion duration. An electrospray voltage of 2.0 kV and capillary temperature of 275°C, with no sheath and auxiliary gas flow, was used.

### Mass Spectrometry Data Analysis

All spectra were analysed using MaxQuant 1.6.2.6 (25), and searched against a combined Homo sapiens and Mtb database. Proteome databases were downloaded from Uniprot on 09 January 2019. The Homo sapiens database contained 42 410 entries while the Mtb

database contained 3 993 entries. Peak list generation was performed within MaxQuant and searches were performed using default parameters and the built-in Andromeda search engine (26). The enzyme specificity was set to consider fully tryptic peptides, and two missed cleavages were allowed. Oxidation of methionine and deamidation of asparagine and glutamine was allowed as variable modification, while carbamidomethylation of cysteine was allowed as a fixed modification. TMT labelling (TMT 10 plex) of peptide N-termini and lysine residues was enabled during the database search. A protein and peptide false discovery rate (FDR) of <1% was employed in MaxQuant. Proteins that contained similar peptides and that could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Reporter ion quantification (TMT 10 plex) at MS3-level was enabled in MaxQuant and reporter ion intensities were used for relative quantification. Reverse hits, contaminants, and proteins only identified by site were removed before further analysis.

Samples were grouped in either end of treatment (EOT) or active tuberculosis (TB) groups in accordance with sampling for this study. The corrected reporter ion intensities were log2 transformed and centred around zero by median normalization. Protein groups were further filtered to contain at least two unique peptides and to contain intensities for all the TMT-10 plex reporter ions before downstream exploratory, statistical, and bioinformatics analysis. Next, principal component analysis was performed to identify possible clusters in the data not based on any a priori knowledge. One of the EOT cases clustered with the active TB cases, indicating a protein signature similar to active TB. Follow up analysis revealed that this participant showed signs of active TB (modified culture confirmed) following successful anti-TB treatment and subsequently relapsed. Therefore, this EOT case was assigned as a TB case during statistical analysis. Hypothesis testing was performed using the Limma package (27), available in the R/Bioconductor repository (28), and p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method (27,28). Proteins were considered significantly regulated when they had at least a two-fold change difference between groups and a corrected p-value <0.05. Enrichment analysis was performed using ClusterProfiler (29) and ReactomePA (30), and Gene Set Enrichment analysis (GSEA) was performed using WebGestalt (31).

## RESULTS

### Differentially Culturable Mtb Are Present at the End of Treatment

There were 22 clinically cured patients (negative sputum MGIT cultures after 20 and 24 weeks of anti-tuberculous treatment) analysed for FDG PET-CT activity recruited

for assessment of DCTB in sputum and BALF (Table 1). Three patients (13.6%) had a positive DCTB status with one of the DCTB positive patients relapsing within the first year following treatment completion. FDG PET-CT scans indicated three main response patterns at the end of treatment: A "resolved" PET response, seen in six (27.3%) patient scans, showing no residual increase in FDG uptake at the end of treatment. An "improved response" type shows residual uptake (but decreased in comparison to diagnosis) in 11 patients (50%). A "mixed response" (characterized by either new lesions or increased uptake in some lung lesions) in five patients (22.7%), including two of the three patients with DCTB (Table 1). Images of the M6 FDG PET-CT scan and chest X-rays at diagnosis (Dx), M6 and M18 follow up of the patients with DCTB are shown in Figure 2. In addition to the set of 22 patients with sputum and BALF samples,

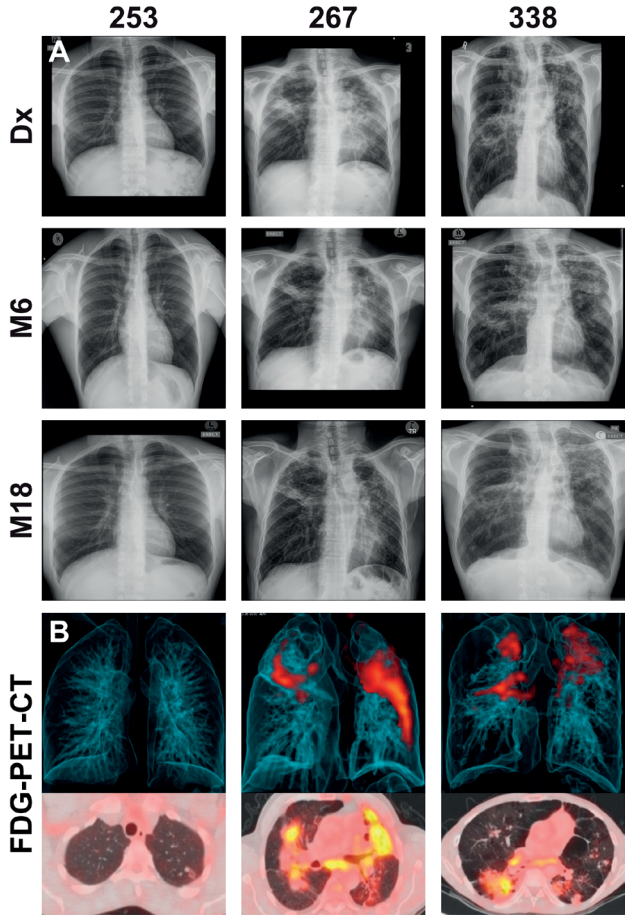



**Figure 2: (A)** Representative chest X-rays at diagnosis (Dx) and month 6 (M6) and **(B)** corresponding FDG-PET CT scan at M6 for cured patients with differentially culturable Mtb. The FDG-PET CT scans show 3-dimensional anterior (top panel) and transverse (lower panels) views.

**Table 1:** Summary of clinical and microbiological parameters of pulmonary TB patients (PID) assessed after successful anti-TB therapy (M6) using the re-suscitation assay (resusc). Age in years, median (range), diagnosis (Dx), body mass index (BMI), chest x-ray (CXR), time to positivity (TTP), time to negativity (TTN).

| PID | Age | Sex | HIV | Hba1c | Smoker | Previous TB | BMI Dx | BMI M6 | CXR M6 | CXR M18 | Culture Dx TTP (days) | Culture TTN (weeks) | Resusc | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 250 | 43 | M | NEG | 6,1 | Daily | NO | 16,8 | 17,8 | Improved | Improved | 8 | 8 | | Cured |
| 253 | 29 | M | NEG | 5,9 | Daily | NO | 19,2 | 21,3 | Worse | Improved | 11 | 16 | YES | Cured |
| 267 | 57 | M | NEG | 6 | EX | YES | 18,2 | 21 | No change | No change | 6 | 16 | YES | Cured |
| 273 | 32 | M | POS | 5,7 | Daily | NO | 18,1 | NA | Improved | Improved | 7 | 8 | | Cured |
| 278 | 36 | M | NEG | 5,4 | NO | YES | 20,8 | 21,6 | Improved | No change | 7 | 8 | | Cured |
| 291 | 33 | M | NEG | 5 | Daily | NO | 16,6 | 19,4 | No change | Improved | 5 | 24 | | Cured |
| 300 | 40 | M | NEG | 5,3 | Daily | NO | 20 | 20,2 | Improved | No change | 4 | 2 | | Cured |
| 306 | 55 | M | NEG | 5 | Daily | NO | 20,9 | 21,2 | No change | Improved | 5 | 8 | | Cured |
| 314 | 42 | F | NEG | 5 | Daily | NO | 22,6 | 24,2 | Improved | No change | 16 | 8 | | Cured |
| 315 | 26 | F | NEG | 5,6 | Daily | NO | 17 | 17,4 | Improved | No change | 4 | 24 | | Cured |
| 318 | 31 | F | NEG | 5,9 | NO | NO | 21,5 | NA | Improved | Improved | 16 | 16 | | Cured |
| 332 | 58 | M | NEG | 6,1 | Daily | NO | 21 | 23,9 | Improved | NA | 5 | 8 | | Cured |
| 334 | 39 | M | NEG | 5,9 | EX | YES | 16,1 | 16,3 | Improved | Worse | 9 | 4 | | Cured |
| 335 | 49 | M | NEG | 5,9 | Daily | YES | 15,4 | 17,2 | Improved | No change | 5 | 8 | | Cured |
| 338 | 48 | M | NEG | 5,3 | Daily | YES | 16 | 16,5 | No change | No change | NA | 8 | YES | Relapse |
| 351 | 22 | F | NEG | 5,5 | Daily | NO | 17,9 | 20,9 | Improved | Improved | 5 | 8 | | Cured |
| 354 | 62 | F | NEG | 6 | Daily | YES | 27,4 | 29,8 | Improved | Improved | 6 | NA | | Cured |
| 357 | 51 | M | NEG | 5,2 | Daily | YES | 18,7 | 18,4 | Improved | No change | NA | NA | | Cured |
| 359 | 48 | F | NEG | 5,5 | Daily | YES | 22,1 | 23 | Improved | No change | 16 | 2 | | Cured |
| 365 | 65 | M | NEG | 5,2 | < daily | YES | 20,6 | 20,5 | Improved | No change | NA | NA | | Cured |
| 370 | 26 | F | NEG | 5,2 | < daily | NO | 16,5 | 18,6 | Improved | No change | 4 | 8 | YES | Cured |
| 375 | 54 | M | NEG | 6,3 | Daily | NO | 15,2 | 17,8 | Improved | Improved | 10 | 8 | | Cured |

**Table 1:** Summary of clinical and microbiological parameters of pulmonary TB patients (PID) assessed after successful anti-TB therapy (M6) using the re-suscitation assay (resusc). Age in years, median (range), diagnosis (Dx), body mass index (BMI), chest x-ray (CXR), time to positivity (TTP), time to negativity (TTN). (*continued*)

| PID | Age | Sex | HIV | Hba1c | Smoker | Previous TB | BMI Dx | BMI M6 | CXR M6 | CXR M18 | Culture Dx TTP (days) | Culture TTN (weeks) | Resusc | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 382 | 43 | F | NEG | NA | Daily | NO | 17,2 | 18,7 | Improved | Improved | 4 | 16 | | Cured |
| 392 | 21 | M | NEG | NA | Daily | NO | 17,5 | 17,9 | Improved | No change | 11 | 8 | | Cured |
| 393 | 39 | M | NEG | NA | Daily | NO | 16,3 | 17,9 | Improved | Improved | NA | 16 | | Cured |
| 394 | 44 | F | NEG | NA | Daily | YES | 25,6 | 26,8 | Improved | Improved | NA | NA | | Cured |
| 396 | 52 | M | NEG | NA | Daily | NO | 20,1 | 22,5 | No change | Improved | 5 | 8 | | Cured |
| 397 | 47 | M | NEG | NA | Daily | YES | 20,3 | 19,6 | No change | Improved | NA | NA | | Cured |
| 404 | 49 | F | NEG | NA | NO | YES | 16,6 | 15,6 | No change | No change | 8 | 24 | YES | Relapse |
| 413 | 29 | M | NEG | NA | Daily | NO | 18,5 | NA | Improved | Improved | 8 | 4 | | Cured |
| 426 | 33 | F | NEG | NA | NO | NO | 22 | NA | Improved | Improved | 9 | 8 | | Cured |
| 425 | 30 | M | NEG | NA | Daily | NO | 19,2 | NA | Improved | Improved | 13 | 4 | | Cured |
| 422 | 57 | M | NEG | NA | Daily | NO | 23,3 | NA | Improved | Improved | 16 | 2 | | Cured |
| 427 | 51 | M | NEG | NA | Daily | NO | 17,2 | NA | Improved | Improved | 9 | 4 | | Cured |
| 428 | 32 | M | NEG | NA | Daily | NO | 15 | NA | Improved | Improved | 7 | 8 | | Cured |
| 429 | 52 | M | NEG | NA | Daily | YES | 15,9 | NA | Improved | Improved | 6 | 8 | | Cured |
| 430 | 45 | M | NEG | NA | Daily | NO | 18,6 | NA | Improved | Improved | 13 | 8 | | Cured |
| 431 | 23 | F | NEG | NA | < daily | NO | 20,4 | NA | Improved | Improved | 5 | 16 | | Cured |
| 390 | 54 | F | NEG | NA | NO | NO | 21,4 | 19,3 | Improved | Improved | 11 | 8 | | Cured |
| 437 | 23 | F | NEG | NA | Daily | NO | 20,8 | NA | Improved | Improved | 5 | 4 | | Cured |
| 433 | 31 | M | NEG | NA | Daily | NO | 26,1 | NA | Improved | Improved | 8 | 2 | | Cured |
| | **43 (21-65)** | **27 Male 65.9%** | | | **34 Smoker 82.93%** | **13 Previous TB 31.7%** | | | **80% Improved 20% No change** | | **8 (2-24)** | | **12.19%** | **4.87%** |

a further 19 clinically cured patients were recruited for assessment of DCTB in their induced sputum only, identifying a further two DCTB positive patients with one of these subsequently relapsing within a year of completing therapy (Table 2). The clinical and microbiological data for all 41 EOT patients is presented in Table S1. Treatment adherence was >96% for all patients and 75% of patients reached their first consecutive negative MGIT 8 weeks into treatment.

**Table 2:** Resuscitation results comparing growth in culture filtrate supplemented media (CFSM) and media only, when using three sample types (bronchoalveolar lavage fluid (BALF), post-bronchoscopy (PB) sputum and induced sputum) in 22 EOT patients. Recovered Mtb strain type was determined using spoligotyping. The true positives row indicates how many of the positive cultures were confirmed to be Mtb using confirmatory assays.

| PID | BALF | | PB sputum | | Induced sputum | | ZN stain | BAP | Selecta | Recovered strain type |
|---|---|---|---|---|---|---|---|---|---|---|
| | CFSM | Media | CFSM | Media | CFSM | Media | | | | |
| 250 | - | - | 3/3 | 1/3 | 2/3 | 1/3 | NEG | POS | NEG | NA |
| 253 | 1/3 | - | 2/3 | 3/3 | 3/3 | 1/3 | POS | NEG | NEG | 1 BEIJING |
| 267 | 3/3 | - | 3/3 | 3/3 | 3/3 | 1/3 | POS | NEG | POS | 373 T1 |
| 273 | | | | | 1/3 | 1/3 | NEG | NA | NA | NA |
| 278 | - | - | - | - | - | - | NA | POS | NEG | NA |
| 291 | * | * | 1/3 | 1/3 | 3/3 | 1/3 | NEG | POS | NEG | NA |
| 300 | 1/3 | - | 2/3 | 1/3 | 3/3 | 1/3 | NEG | POS | NEG | NA |
| 306 | - | - | 1/3 | - | 3/3 | 2/3 | NEG | POS | NEG | NA |
| 314 | 1/3 | - | - | 1/3 | - | - | NEG | POS | NEG | NA |
| 315 | 1/3 | 1/3 | 1/3 | 1/3 | 3/3 | 3/3 | NEG | POS | NEG | NA |
| 318 | 1/3 | 1/3 | | 1/3 | 2/3 | 2/3 | NEG | POS | NEG | NA |
| 334 | - | - | - | - | 3/3 | 2/3 | NEG | POS | NEG | NA |
| 332 | - | - | - | - | - | - | NA | NA | NA | NA |
| 335 | 1/3 | 1/3 | 1/3 | 1/3 | 1/3 | 1/3 | NEG | POS | NEG | NA |
| 338 | 3/3 | 1/3 | 2/3 | 3/3 | 1/3 | 2/3 | POS | NEG | POS | 1 BEIJING |
| 351 | - | - | - | - | 1/3 | 2/3 | NEG | POS | NEG | NA |
| 357 | - | - | - | - | - | - | NA | NA | NA | NA |
| 359 | - | - | 1/3 | - | 1/3 | - | NEG | POS | NEG | NA |
| 365 | 1/3 | - | - | 1/3 | 1/3 | 1/3 | NEG | POS | NEG | NA |
| 375 | - | - | - | - | 1/3 | 2/3 | NEG | POS | NEG | NA |
| 382 | - | 1/3 | 2/3 | 1/3 | 1/3 | 1/3 | NEG | POS | NEG | NA |
| 392 | - | - | 1/3 | - | 1/3 | - | NEG | POS | NEG | NA |
| **% positive** | **41** | **22,7** | **54,5** | **54,5** | **81,8** | **72,7** | | | | |
| **% true positive** | **13,6** | **4,5** | **4,5** | **0** | **9** | **0** | | | | |

Growth is indicated as either positive in three of the replicate wells (3/3), two of the three replicates (2/3) or one of the three replicate wells (1/3) or no growth (-). Grey blocks indicate that a sample was not available. * Indicates the wells became contaminated with fungal growth and confirmatory assays could not be conducted. The green blocks indicate the samples that were confirmed positive for *Mtb*.

## BALF Is the Optimal Sample Type to Identify Differentially Culturable Mtb at the End of Treatment

The DCTB assay indicated that a significant portion of culture wells were positive for growth in all samples analysed by visual inspection [induced sputum (81.8%), post-bronchoscopy sputum (54.5%) and BALF (41%)] (Table 3). However, the majority of these did not contain Mtb, as determined by confirmatory assays. Only samples that could be sub-cultured onto selective media and subsequently strain typed were

**Table 3:** Outcomes of patients assessed by FDG PET CT and the modified culture assay (resuscitation) at month 6 after successful anti-TB therapy. PET CT responses were classified according to total glycolytic activity (TGAI), cavity, response pattern and intensity rank. Highlights in red show patients classified as higher risk according to PET CT cut offs (TGAI ≥600, cavity volume ≥7mL, mixed response and very high intensity rank).

| PID | Resuscitation | Outcome | Response Pattern | Intensity rank | TGAI | Total cavity volume (mL) |
|---|---|---|---|---|---|---|
| 338 | YES | Relapse | Mixed | Very high | 43007,0 | 39,9 |
| 267 | YES | Cured | Mixed | Very high | 38304,0 | 30,9 |
| 253 | YES | Cured | Resolved | Minimal | 70,36 | 0,0 |
| Median | | | | | 38304,00 | 30.9 |
| (range) | | | | | (70.4-43007) | (0-39.9) |
| 332 | | Cured | Mixed | High | 4598,2 | 0,0 |
| 351 | | Cured | Mixed | Moderate | 2853,4 | 0,0 |
| 250 | | Cured | Mixed | Very high | 620,0 | 62,9 |
| 334 | | Cured | Improved | Very high | 9757,8 | 19,4 |
| 375 | | Cured | Improved | Very high | 9562,6 | 1,8 |
| 291 | | Cured | Improved | Moderate | 6349,8 | 5,9 |
| 392 | | Cured | Improved | Very high | 6109,6 | 0,0 |
| 382 | | Cured | Improved | Very High | 2795,4 | 2,8 |
| 335 | | Cured | Improved | High | 2265,7 | 15,7 |
| 357 | | Cured | Improved | Moderate | 634,9 | 0,0 |
| 278 | | Cured | Improved | Mild | 585,6 | 0,0 |
| 315 | | Cured | Improved | Moderate | 235,5 | 0,0 |
| 300 | | Cured | Improved | High | 173,2 | 0,0 |
| 314 | | Cured | Improved | Mild | 151,7 | 0,0 |
| 273 | | Cured | Resolved | None | 1681,9 | 0,0 |
| 365 | | Cured | Resolved | Minimal | 951,7 | 0,0 |
| 318 | | Cured | Resolved | Minimal | 306,2 | 0,5 |
| 306 | | Cured | Resolved | Minimal | 213,8 | 0,0 |
| 359 | | Cured | Resolved | Minimal | 168,2 | 2,8 |
| Median | | | | | 793.27 | 0.0 |
| (range) | | | | | (151.7-9757.8) | (0.0-62.9) |
| | | | **22.7% Mixed** | | | |
| | | | **50% Improved** | | | |
| | | | **27.3% Resolved** | | | |

declared positive for DCTB. As shown in Table 3, the number of false positives was highest in the induced sputum and lowest in BALF. The relapse patients (in samples taken before relapse) showed pure Mtb growth, without obvious contamination by commensal bacteria, in all sample types tested (BALF, induced and post-bronchoscopy sputum for PID338 and induced sputum for PID404) (Figure S1) and these grew rapidly in the micro-well format (<5 days). BALF was also the sample in which we could consistently detect all DCTB positive patients when comparing to induced sputum and post-bronchoscopy sputum (Table S1). Differential cell staining showed that sputa is comprised primarily of epithelial cells, whereas the cell profile of BALF samples is comprised of macrophages, lymphocytes, and neutrophils (Figure S2).

## An Altered Host Proteome Can Be Identified at the End of Treatment

Distinct FDG PET-CT responses at the end of treatment prompted us to investigate the host response at the site of disease using mass spectrometry-based proteomics. We analysed BALF supernatant collected from a subset of the patients recruited in this study, namely five EOT patients, with a mixture of FDG PET-CT responses (Table 4) and five pre-treatment active TB cases, using multiplexed proteomics with tandem mass tags (TMT 10 plex). This relative proteome comparison would enable us to gain fundamental insights into the host response at the site of disease during active TB and in clinically cured patients at the end of 6 month standard anti-TB therapy.

We identified a total of 1 565 high confidence proteins in BALF from EOT and active TB cases, which contained more than two unique peptides and were identified at a 1% empirical protein false discovery rate (FDR). These proteins spanned six orders of magnitude in abundance (Figure S3). The most abundant protein was albumin, followed by several other proteins commonly identified in plasma, including serotransferrin, haptoglobin, alpha-1-acid glycoprotein, and alpha-2-macroglobulin. Importantly, we were able to probe the BALF proteome to a level where we could identify cytokines (interleukin-18, macrophage colony-stimulating factor 1 and macrophage inhibitory factor), chemokines (CXCL17), pulmonary surfactants (pulmonary surfactant-associated protein A, B, and D) and antimicrobial peptides (cathelicidin LL37 and neutrophil defensins) in these samples (Figure S3) identifying these low abundance proteins in clinical samples, especially biological fluids, is notoriously difficult. The proteins identified in BALF from EOT and active TB were mainly involved in processes associated with neutrophil activation, innate and adaptive immune responses, RNA metabolism and translation (Figure S4).

For quantitative analysis, the corrected reporter ion intensities for each sample was log 2 transformed, resulting in a normal distribution (Figures S5A–J). The data was further

normalized by subtracting the median to centre all distributions around zero and thereby minimize inter sample variation (Figure S5K). This data was used for further analysis. We were able to reproducibly quantify 1 521 proteins in all patient samples analysed, after filtering out common contaminants such as albumin and keratin.

Principal component analysis was used to identify the major sources of variation between the BALF proteomes of active TB and EOT cases, and to identify clusters in our data driven by protein expression profiles. As hypothesized, a clear separation was observed between the active TB and the EOT samples in the first component (Figure 3A). Interestingly, one EOT sample from this clinically cured set (PID 338) clustered with the active TB samples. During follow-up, this participant was subsequently diagnosed with recurrent TB disease (Figure 3A). Even though this participant completed 6 months of anti-TB treatment and was declared cured, a similar protein expression profile to that of an active TB case was observed 3 months before the participant returned to the clinic and was diagnosed as a relapse. To further confirm that the relapsed patient clusters with the TB cohort, we performed hypothesis testing by grouping the relapse as either active TB, EOT cure or removed from the sample set. We found that when relapse is added to the TB group there is an increase of 38 additional proteins (Figure S6) that distinguish active disease from cure. However, addition of the relapse case as an EOT cure resulted in little to no differential abundance of proteins due to the effects of the standard deviation which were introduced due to the active TB profile of the relapse distorting the EOT cured profiles. We therefore treated this sample as an active TB case for downstream analysis. We used hypothesis testing to determine the differentially regulated proteins between active TB and EOT cases. In total, 266 proteins were differentially regulated (two-fold change and adjusted p-value <0.05), where 138 proteins were significantly down regulated and 128 proteins were significantly up regulated during active TB disease (Figure 3B, Supplementary file 1). In the active TB cases, the top 15 regulated proteins by fold change included inflammatory and immune proteins such as CRP, BPIA2, and A1AG1 (ORM1) (Figure 3C). Furthermore, active TB was associated with increased levels of immunoglobulins, proteins associated with infiltration of neutrophils (MMP8, DEFA3, PRTN3), complement cascade members (C1QB, C1R, C1RL, C1S, C3, C4A, C5, C8A), and proteins playing a role in blood coagulation (F5, F13B, VWF, FGG, FGB, GP1BA). The top 15 down regulated proteins in the active TB cases were associated with translation as indicated by the presence of multiple ribosomal proteins (Figure 3D). To gain insights into the processes regulated between patients with active TB compared to those at the end of treatment (EOT), we performed gene set enrichment analysis on the differentially regulated proteins using non-redundant gene ontology terms. The greatest enrichment ratios were associated with proteins with increased abundance in the active TB cases, of which the humoral

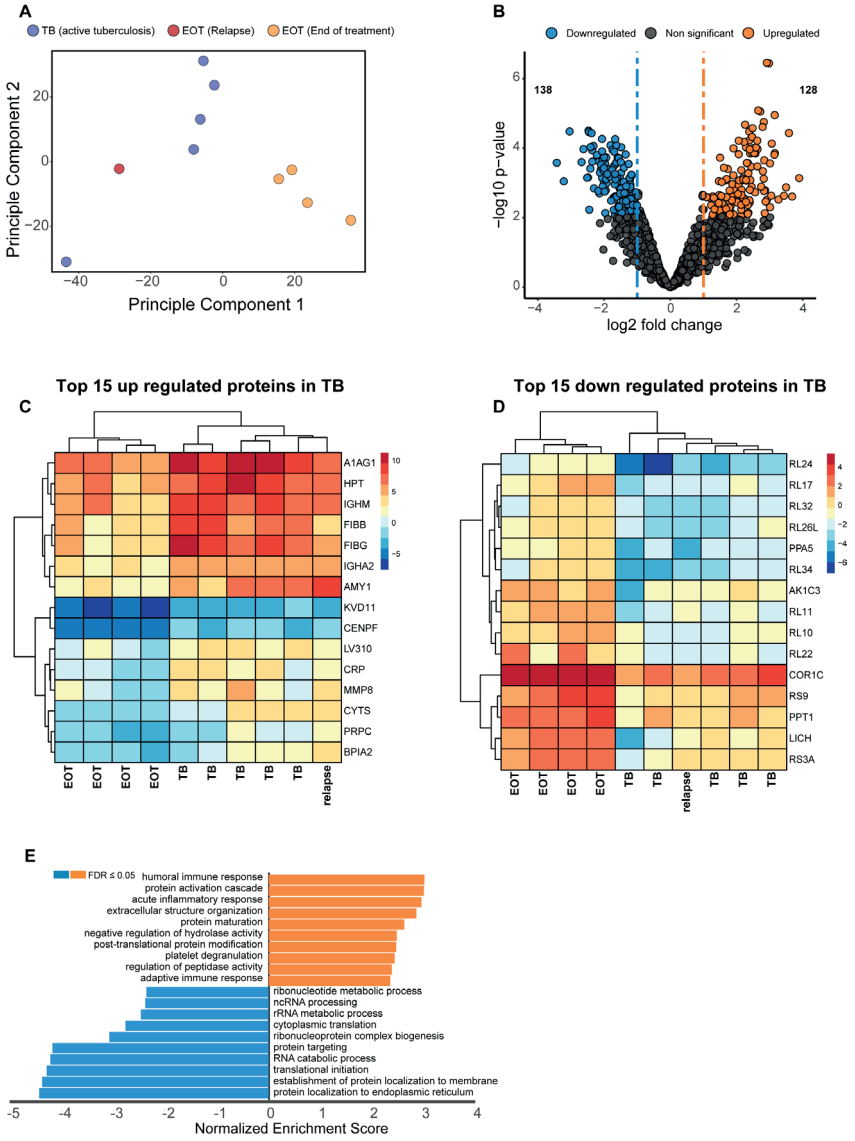



**Figure 3:** Unique proteome signatures exist in BALF from patients with active TB compared to patients with clinical cure. **A)** Principle component analysis of protein intensities obtained by mass spectrometry analysis of BALF obtained from active TB patients at baseline (TB) and patients successfully completing 6 -months anti -TB therapy (EOT). Active TB cases could be distinguished from the majority of EOT cases in principle component 1, with the exception of one EOT case. Follow up analysis of this patient revealed this was the relapse case with differentially culturable M. tuberculosis. **B)** Volcano plot representing the differentially regulated proteins in active TB compared to EOT. The relapse sample was considered as a TB case for the statistical test. Heatmaps of the **C)** top 15 upregulated and **D)** downregulated proteins from the hypothesis testing. **E)** Gene enrichment analysis (GSEA) of all significantly regulated proteins against gene ontology terms. GSEA was performed using the WebGestalt webserver.

response and acute inflammatory response featured prominently (Figure 3E). On the other hand, proteins associated with RNA metabolic processes, protein synthesis and translation were more abundant in EOT cases.

## DISCUSSION

Assessing true sterilization at the end of anti-TB therapy is paramount to reducing the global TB burden, especially since one of the highest priorities of the TB control programs is shortening treatment (1). A key area of contention is the presence or absence of "persistent" Mtb cells in patients and whether this quiescent population is responsible for subsequent relapse after therapy has ceased (32,33). Here, we have shown that although patients at the end of successful anti-TB therapy are cured by clinical standards, we were able to identify a fraction of patients retaining DCTB at the end of treatment, two of which subsequently relapsed within the first year after stopping therapy. While the small sample size in this exploratory study does not have any statistical power, it is notable that Mtb growth was confirmed in CFSM from induced sputum of the two cases with the highest residual inflammatory burden on EOT FDG PET-CT, including the patient with subsequent relapse. This suggests that a correlation between the presence of DCTB and clinical outcome is plausible and that it should be further explored in prospective studies.



We decided to investigate both sputa and site of disease (BALF) for the presence of DCTB, as patients cannot always produce a sputum at the end of treatment. Using the DCTB assay at the end of treatment, most sputum sample types showed growth of commensal bacteria, even with routine decontamination procedures. Unsurprisingly, BALF was the optimal sample type for recovering differentially culturable bacteria, since BALF represents a direct representation of the site of disease, and with less exposure to commensal organisms. This has implications for the routine use of the DCTB assay for monitoring sterilizing cure from sputum in a clinical setting. While the lengthy and laborious nature of the DCTB assay would make large-scale implementation of the technique impractical, these results indicate that these DCTB cells are an important consideration when monitoring treatment response, specifically for treatment shortening studies. Viability assays measuring Mtb rRNA as a marker of treatment response currently show the most promise as alternatives to measure bacterial viability in patient samples (34), and these should be tested in parallel to determine their ability to identify these differentially culturable bacilli.

The relapse rate observed in this study is consistent with rates seen in other studies, although we acknowledge the small number of cases (35–38). Even though we cannot define the physiological state of the mycobacterial cells before recovery, this data shows that there remains a sub-population of cells that are undetectable by standard culture (MGIT), yet differentially culturable in the presence of CFSM. These results build on previous studies that have shown that DCTB can be detected in patient sputa from TB patients at baseline and early chemotherapy (12,13,39) and that bacterial persistence is likely the cause of EOT inflammation (4). The fact that many more patients retain FDG avid lesions at EOT than have demonstrable DCTB may be due to a low number of viable bacteria or relative insensitivity of the DCTB assay. Further work is needed to examine the link between inflammation on PET-CT imaging and DCTB. Multiple studies have been necessary to examine this question especially when trying to correlate with relapse since only 4–10% of patients relapse. This study confirms and extends the previous studies by modifying the DCTB assay and by thoroughly examining the patients (PET-CT, BALF proteomics) specifically at the end of successful treatment. To our knowledge, this is the first study to investigate the presence of differentially culturable Mtb in sputum and BALF at the end of anti-TB therapy.

The patients that were identified as positive for DCTB yet have maintained cure indicate how the host immune response plays a significant role to curtail the remaining infection and direct treatment outcome. Indeed, although the factors responsible for relapse and reactivation have not yet been completely elucidated, immune suppression has a clear impact on this phenomenon (9,40–42). As we have shown previously, a large proportion of patients at the end of successful therapy show ongoing inflammation as detected by FDG PET-CT indicating patterns consistent with active disease (4), yet not all these patients will go on to relapse. In this study, we found that FDG PET-CT responses consistent with active TB only partially correlated with differentially culturable Mtb. One patient was positive for DCTB, yet FDG PET-CT indicated complete resolution of lesions. It must be noted that this patient sample was only positive for Mtb in the BALF sample in one of the triplicate wells, unlike the other patients, where Mtb was consistently detected in all samples analysed. It would also be beneficial to assess how long these DCTB cells can be detected, as it is possible that although they are present at month 6, later time points would show that these have been cleared by the immune response. The heterogeneity in patient FDG PET-CT responses highlights how different lesions, even within the same host, can progress differently. It remains to be determined what factors determine whether a patient achieves true sterilization.

We therefore analysed the global host proteome profiles of five end of treatment patients (consisting of one EOT patient that subsequently relapsed and four that had maintained

cure with a mixture of FDG PET-CT responses) vs. active TB patients (before treatment) at the site of disease (BALF), to enhance our understanding of the lung microenvironment. Principal component analysis indicated distinct proteomic signatures stratifying active and clinically cured TB patients into distinct groups. Additionally, the relapsed participant, although standard sputum MGIT negative and clinically defined as cured at M6, produced a M6 BALF protein expression profile that clustered more closely with the active TB group than with the clinically cured group. Taken together, it is likely that patients who are prone to relapse can be detected within the BAL fluid, however further investigation is required with greater study cohorts.

Gene set enrichment analysis revealed that the major upregulated responses in active TB (pre-treatment samples) correspond to important immune responses notably, the acute inflammatory response and protein activation cascades. The acute phase response serves various important functions within the immune system, one of which is the activation of the complement cascade by positive acute-phase proteins (APPs) (43). Several APPs that are involved in the complement cascade were identified as upregulated in the BALF of active TB cases including CRP, C3, C5, and C1qrs pointing to classical and lytic activation pathways. The complement system is a key component of the innate immune response and is initiated rapidly upon infection to promote inflammation, recruit leukocytes, form a membrane attack complex and ultimately destroy invading pathogens (44). Although there are limited BALF proteomics studies in the context of TB, a previous report showed that BALF from healthy patients contains classical complement activity which leads to C3b binding of Mtb *in vitro* (45). Our findings are also consistent with proteomic investigations of blood-based signatures for TB progression, showing a similar elevation of members of the complement cascade, infiltration of neutrophils and blood coagulation during early disease (46–48), whilst providing additional information on altered pathways of treatment response. It stands to reason that following treatment and sterilizing cure, these pathways would return to lower levels. Unsurprisingly, we observed these processes in actively infected TB patients and their decrease in the M6 EOT clinically cured patients. Further characterization in a larger cohort of patients would be beneficial in assessing whether these proteomic signatures are consistent, as these results show promise in identifying distinguishing signatures of treatment response. If these observations can be extended, emphasis should be placed on replacing the invasive BALF procedure with a more easily obtained sample. Analysis of serum samples by multiplexed ion mobility spectrometry similarly shows a similar decrease of acute phase proteins in response to antibiotic treatment and culture status (49). This supports the clinical translatability of identified BALF markers being applied for blood-based testing. To our knowledge, this is the first study to investigate host proteome changes at the site of disease during active TB and follow-

ing anti-TB therapy. We were able to identify a large number of proteins spanning a broad dynamic range offering new avenues for targeted investigations.

## CONCLUSION

This study, although exploratory in nature, shows a comprehensive characterization of end of treatment patients from both a host and bacterial angle. We describe that DCTB can be detected at the end of successful anti-TB therapy and may represent a risk for relapse. Our methods were labour intensive and because of this they were limited in scope. Targeting dormant bacilli has been suggested as an important avenue to fight re-activation of TB and shorten treatment (50), and this study defines a pathway to further explore biomarkers of persistence in patients at the end of therapy.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary Material.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Human Research Ethics Committee of Stellenbosch University (N10/01/013 and N16/05/070). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

**CB, GW, SM, JW**, and **TH:** study concept and design. **SM**: patient recruitment and visual scan analysis. **TH, JG**, and **CB**: acquisition, analysis, and interpretation of the proteomics data. **CB** and **RV**: acquisition of the microbiological data. **CB**: drafting of manuscript. **CB, GW, TH, JG, SM, JW, RV, BK, NP, MT**, and **AL**: manuscript revision and important intellectual content. All authors contributed to the article and approved the submitted version.

## FUNDING

The financial assistance of the National Research Foundation (NRF) toward this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author(s) and are not necessarily to be attributed to the NRF. Research reported in this publication was supported by the South African Medical Research Council. The content is solely the responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council. Funding was provided by the Catalysis Foundation for Health through the BMGF (OPP51919). GW leads the South African Research Chair Initiative (SARChI) in TB Biomarkers (#86535). RV acknowledges funding from the National Research Foundation (NRF), South African Medical Council (SAMRC), and Stellenbosch University.

## CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



## SUPPLEMENTARY DATA

Supplementary data for not in this document can be accessed at the following URL with a Mendeley account:

https://tinyurl.com/5u3sumvm

## REFERENCES


1. Organization WH. Global tuberculosis report 2019 [Internet]. Geneva PP - Geneva: World Health Organization; Available from: https://apps.who.int/iris/handle/10665/329368
2. Naidoo K, Dookie N. Insights into Recurrent Tuberculosis: Relapse Versus Reinfection and Related Risk Factors. In: Tuberculosis [Internet]. InTech; 2018 [cited 2020 Nov 27]. p. 13. Available from: http://dx.doi.org/10.5772/intechopen.73601
3. Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, Gey van Pittius NC, et al. Patients with active tuberculosis often have different strains in the same sputum specimen. Am J Respir Crit Care Med. 2004 Mar;169(5):610–4.
4. Malherbe STST, Shenai S, Ronacher K, Loxton AGAG, Dolganov G, Kriel M, et al. Persisting positron emission tomography lesion activity and Mycobacterium tuberculosis mRNA after tuberculosis cure. Nat Med. 2016 Oct;22(10):1094–100.
5. Ambreen A, Jamil M, Rahman MA ur, Mustafa T. Viable Mycobacterium tuberculosis in sputum after pulmonary tuberculosis cure. BMC Infect Dis [Internet]. 2019;19(1):923. Available from: https://doi.org/10.1186/s12879-019-4561-7
6. Balaban NQ, Merrin J, Chait R, Kowalik L, Leibler S. Bacterial Persistence as a Phenotypic Switch. Science (80- ) [Internet]. 2004 Sep 10;305(5690):1622 LP – 1625. Available from: http://science.sciencemag.org/content/305/5690/1622.abstract
7. Gefen O, Balaban NQ. The importance of being persistent: heterogeneity of bacterial populations under antibiotic stress. FEMS Microbiol Rev. 2009 Jul;33(4):704–17.
8. Lewis K. Persister Cells. Annu Rev Microbiol [Internet]. 2010 Sep 30;64(1):357–72. Available from: https://doi.org/10.1146/annurev.micro.112408.134306
9. Manina G, Dhar N, McKinney JD. Stress and host immunity amplify Mycobacterium tuberculosis phenotypic heterogeneity and induce nongrowing metabolically active forms. Cell Host Microbe. 2015 Jan;17(1):32–46.
10. Walter ND, Dolganov GM, Garcia BJ, Worodria W, Andama A, Musisi E, et al. Transcriptional Adaptation of Drug-tolerant Mycobacterium tuberculosis During Treatment of Human Tuberculosis. J Infect Dis. 2015 Sep;212(6):990–8.
11. Fisher RA, Gollan B, Helaine S. Persistent bacterial infections and persister cells. Nat Rev Microbiol [Internet]. 2017;15(8):453–64. Available from: https://doi.org/10.1038/nrmicro.2017.42
12. Mukamolova G V., Turapov O, Malkin J, Woltmann G, Barer MR. Resuscitation-promoting factors reveal an occult population of tubercle bacilli in sputum. Am J Respir Crit Care Med. 2010;181(2):174–80.
13. Chengalroyen MD, Beukes GM, Gordhan BG, Streicher EM, Churchyard G, Hafner R, et al. Detection and Quantification of Differentially Culturable Tubercle Bacteria in Sputum from Patients with Tuberculosis. Am J Respir Crit Care Med. 2016 Dec;194(12):1532–40.
14. Russell DG, Cardona P-JJ, Kim M-JJ, Allain S, Altare FF. Foamy macrophages and the progression of the human tuberculosis granuloma. Nat Immunol [Internet]. 2009 Sep;10(9):943–8. Available from: http://dx.doi.org/10.1038/ni.1781
15. Gengenbacher M, Kaufmann SHE. Mycobacterium tuberculosis : Success through dormancy. FEMS Microbiol Rev. 2013;36(3):514–32.
16. Ehlers S, Schaible UE. The granuloma in tuberculosis : dynamics of a host – pathogen collusion. 2013;3(January):1–9.

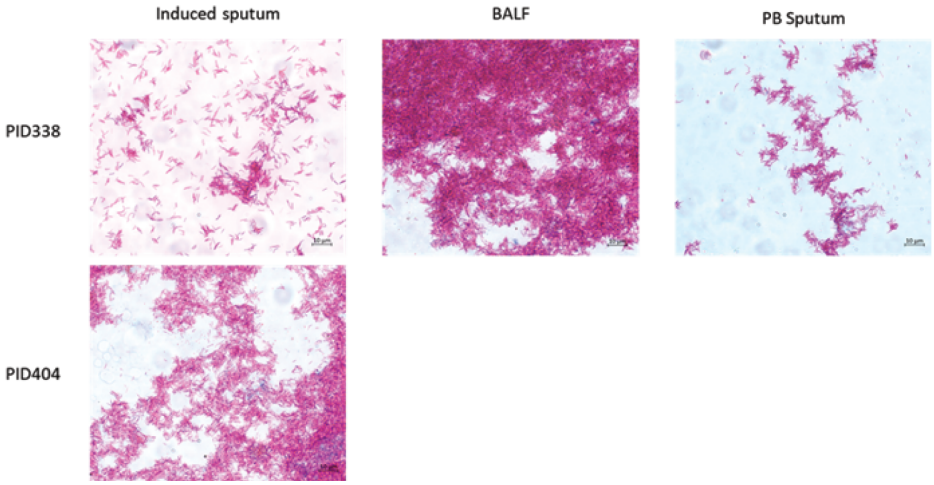
17. Gollan B, Grabe G, Michaux C, Helaine S. Bacterial Persisters and Infection: Past, Present, and Progressing. Annu Rev Microbiol. 2019;73(1):359–85.
18. McCune RM, McDermott W, Tompsett R. The fate of Mycobacterium tuberculosis in mouse tissues as determined by the microbial enumeration technique. II. The conversion of tuberculous infection to the latent state by the administration of pyrazinamide and a companion drug. J Exp Med. 1956;104(5):763–802.
19. McCune RM, Feldmann FM, McDermott W. Microbial persistence. II. Characteristics of the sterile state of tubercle bacilli. J Exp Med. 1966;123(3):469–86.
20. Griffin AJ, Li L-X, Voedisch S, Pabst O, McSorley SJ. Dissemination of persistent intestinal bacteria via the mesenteric lymph nodes causes typhoid relapse. Infect Immun. 2011/01/24. 2011 Apr;79(4):1479–88.
21. Kaiser P, Regoes RR, Dolowschiak T, Wotzka SY, Lengefeld J, Slack E, et al. Cecum Lymph Node Dendritic Cells Harbor Slow-Growing Bacteria Phenotypically Tolerant to Antibiotic Treatment. PLOS Biol. 2014 Feb;12(2):e1001793.
22. Loraine J, Pu F, Turapov O, Mukamolova G V. Development of an In Vitro Assay for Detection of Drug-Induced Resuscitation-Promoting-Factor-Dependent Mycobacteria. Antimicrob Agents Chemother. 2016 Oct;60(10):6227–33.
23. Streicher EM, Victor TC, van der SG, Sola C, Rastogi N, van Helden PD, et al. Spoligotype signatures in the Mycobacterium tuberculosis complex. J Clin Microbiol [Internet]. 2007;45. Available from: http://dx.doi.org/10.1128/JCM.01429-06
24. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods [Internet]. 2009 [cited 2020 Nov 24];6(5):359–62. Available from: https://pubmed.ncbi.nlm.nih.gov/19377485/
25. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
26. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V., Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805.
27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Apr 20 [cited 2018 Mar 15];43(7):e47–e47. Available from: http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for
28. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
29. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012/03/28. 2012 May;16(5):284–7.
30. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12(2):477–9.
31. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res [Internet]. 2019 [cited 2020 May 26];47:199–205. Available from: https://academic.oup.com/nar/article-abstract/47/W1/W199/5494758
32. Zhang Y. Persisters, persistent infections and the Yin-Yang model. Emerg Microbes Infect. 2014;3.

33. Mandal S, Njikan S, Kumar A, Early J V., Parish T. The relevance of persisters in tuberculosis drug discovery. Microbiol (United Kingdom). 2019;165(5):492–9.
34. Honeyborne I, McHugh TD, Phillips PPJ, Bannoo S, Bateson A, Carroll N, et al. Molecular bacterial load assay, a culture-free biomarker for rapid and accurate quantification of sputum Mycobacterium tuberculosis bacillary load during treatment. J Clin Microbiol. 2011;49(11):3905–11.
35. Sonnenberg P, Murray J, Glynn JR, Shearer S, Kambashi B, Godfrey-Faussett P. HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: A cohort study in South African mineworkers. Lancet. 2001;358(9294):1687–93.
36. Luzze H, Johnson DF, Dickman K, Mayanja-Kizza H, Okwera A, Eisenach K, et al. Relapse more common than reinfection in recurrent tuberculosis 1-2 years post treatment in urban Uganda. Int J Tuberc Lung Dis [Internet]. 2013 Mar 1 [cited 2020 Oct 30];17(3):361–7. Available from: /pmc/articles/PMC6623981/?report=abstract
37. Gillespie SH, Crook AM, McHugh TD, Mendel CM, Meredith SK, Murray SR, et al. Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. N Engl J Med [Internet]. 2014 Oct 23 [cited 2020 Oct 30];371(17):1577–87. Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa1407426
38. Merle CS, Fielding K, Sow OB, Gninafon M, Lo MB, Mthiyane T, et al. A Four-Month Gatifloxacin-Containing Regimen for Treating Tuberculosis. N Engl J Med. 2014;371(17):1588–98.
39. Dusthackeer A, Balasubramanian M, Shanmugam G, Priya S, Nirmal CR, Sam Ebenezer R, et al. Differential Culturability of Mycobacterium tuberculosis in Culture-Negative Sputum of Patients With Pulmonary Tuberculosis and in a Simulated Model of Dormancy . Vol. 10, Frontiers in Microbiology . 2019. p. 2381.
40. Connolly LE, Edelstein PH, Ramakrishnan L. Why is long-term therapy required to cure tuberculosis? PLoS Med. 2007;4(3):435–42.
41. Gideon HP, Flynn JL. Latent tuberculosis: what the host "sees"? Immunol Res. 2011;50(2–3):202–12.
42. Esmail H, Barry CE, Young DB, Wilkinson RJ. The ongoing challenge of latent tuberculosis. Philos Trans R Soc B Biol Sci. 2014;369(1645):20130437–20130437.
43. Jain S, Gautam V, Naseem S. Acute-phase proteins: As diagnostic tool. J Pharm Bioallied Sci. 2011 Jan;3(1):118–27.
44. Ricklin D, Hajishengallis G, Yang K, Lambris JD. Complement: a key system for immune surveillance and homeostasis. Nat Immunol. 2010/08/19. 2010 Sep;11(9):785–97.
45. Ferguson JS, Weis JJ, Martin JL, Schlesinger LS. Complement Protein C3 Binding to Mycobacterium tuberculosis Is Initiated by the Classical Pathway in Human Bronchoalveolar Lavage Fluid. Infect Immun. 2004;72(5):2564–73.
46. Scriba TJ, Penn-Nicholson A, Shankar S, Hraha T, Thompson EG, Sterling D, et al. Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease. PLoS Pathog. 2017 Nov;13(11):e1006687.
47. Esmail H, Lai RP, Lesosky M, Wilkinson KA, Graham CM, Horswell S. Complement pathway gene activation and rising circulating immune complexes characterize early disease in HIV-associated tuberculosis. 2018;
48. Penn-Nicholson A, Hraha T, Thompson EG, Sterling D, Mbandi SK, Wall KM, et al. Discovery and validation of a prognostic proteomic signature for tuberculosis progression: A prospective cohort study. PLOS Med [Internet]. 2019 Apr 16;16(4):e1002781. Available from: https://doi.org/10.1371/journal.pmed.1002781

49. Kedia K, Wendler JP, Baker ES, Burnum-Johnson KE, Jarsberg LG, Stratton KG, et al. Application of multiplexed ion mobility spectrometry towards the identification of host protein signatures of treatment effect in pulmonary tuberculosis. Tuberculosis (Edinb). 2018/07/18. 2018 Sep;112:52–61.
50. Fattorini L, Piccaro G, Mustazzolu A, Giannoni F. Targeting Dormant Bacilli To Fight Tuberculosis. Mediterr J Hematol Infect Dis. 2013;5(1):2013072.

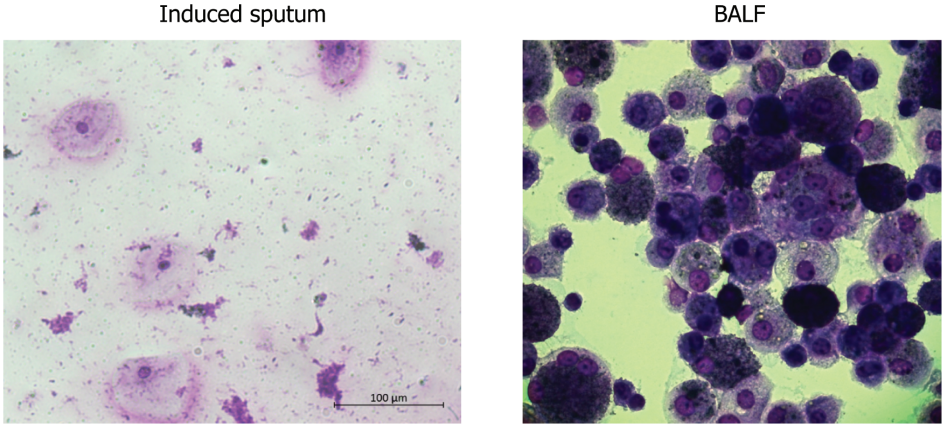
# SUPPLEMENTARY FIGURES

## Supplementary figure 1



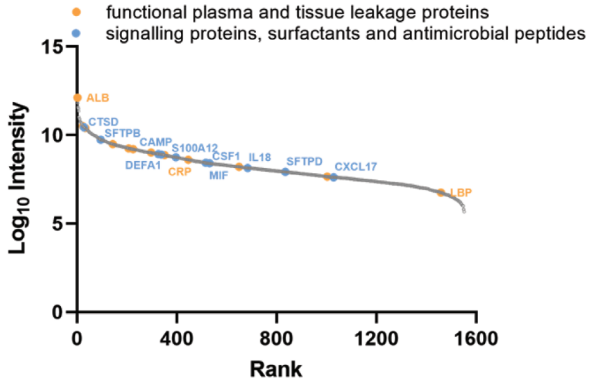

**Supplementary figure 1**: Ziehl-Neelsen (ZN) stain of acid-fast bacilli showing purity of the recovered culture using the resuscitation assay in induced sputum (PID 338 and 404) and bronchoalveolar lavage fluid (BALF) and post-bronchoscopy sputum (PID338).

## Supplementary figure 2



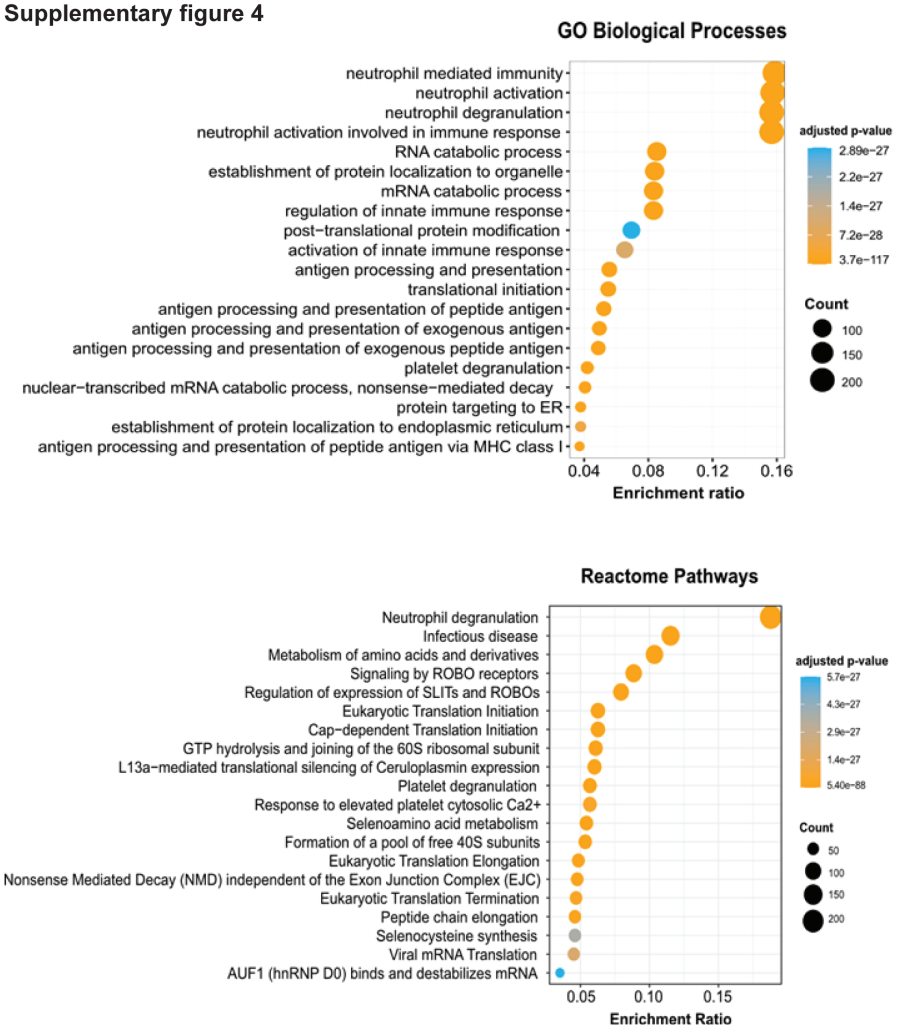

**Supplementary figure 2**: Representative microscopy images of differential cell staining of an induced sputum (100X) and BALF sample (40X) from the same EOT patient. Samples were concentrated by cytocentrifugation at 95rpm for 7 minutes with slow acceleration and stained using Rapid-Diff stain.

**Supplementary figure 3**



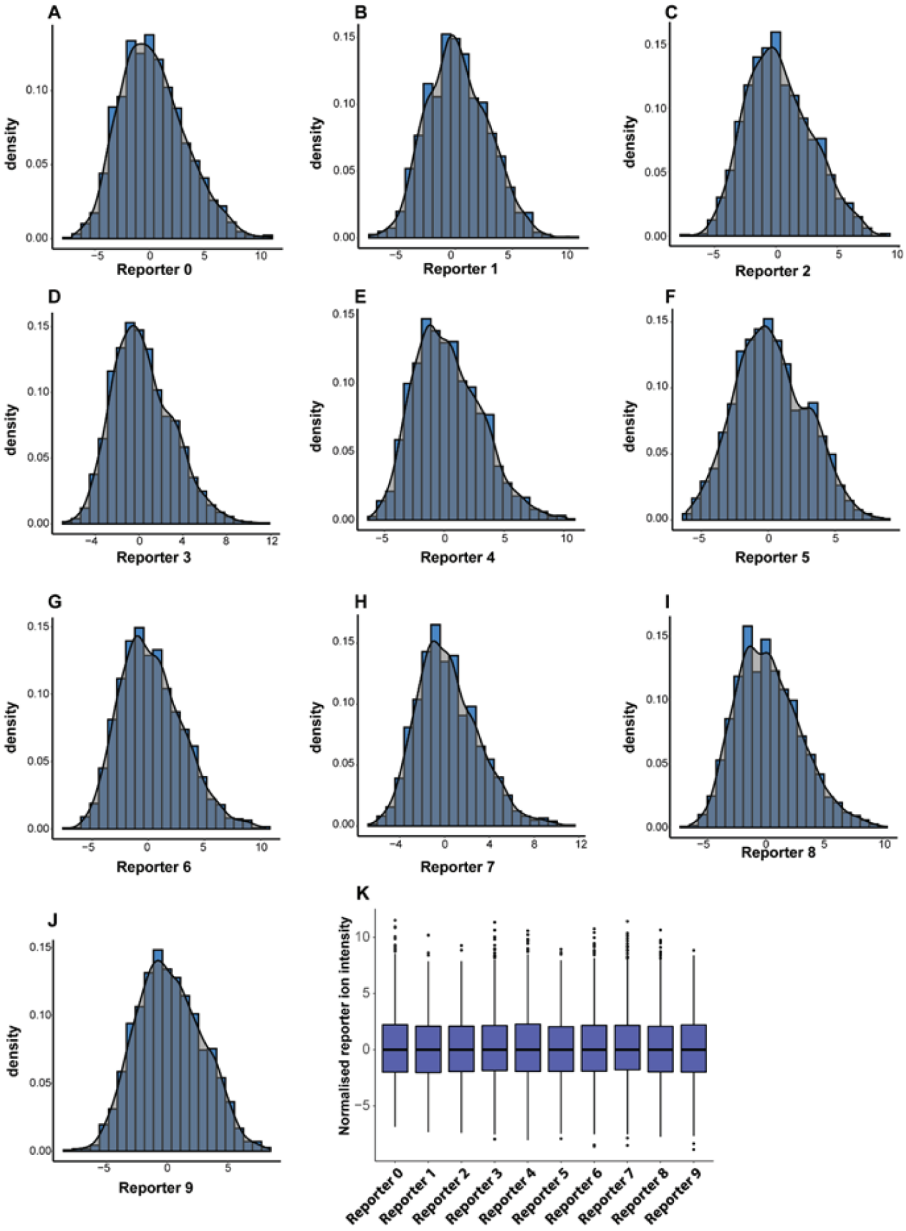

**Supplementary figure 3:** Dynamic range estimation showing combined log10 intensity of the proteins identified in the study with functional plasma and tissue leakage proteins shown in orange and signaling proteins, surfactants and antimicrobial peptides shown in blue. The plot shows a dynamic range of 6 orders of magnitude.

**Supplementary figure 4**



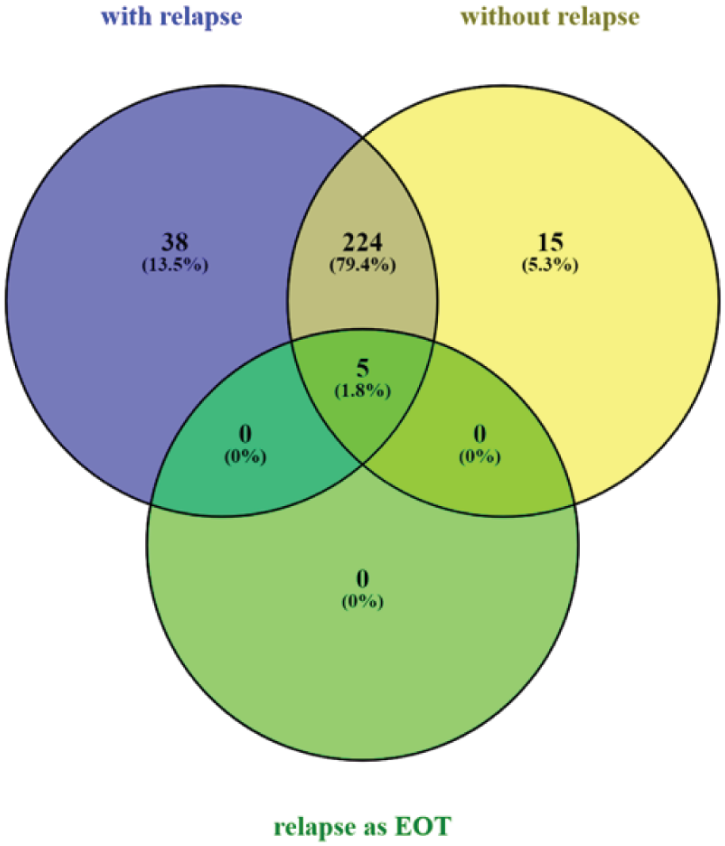

**Supplementary figure 4.** Dotplot showing enrichment results of the GO biological processes and Reactome pathways. The size of the circle signifies the count of the biological process or reactome pathway. The colour denotes the adjusted p-value.

## Supplementary figure 5




**Supplementary figure 5.** Quality control of reporter ion intensities after normalization. **A-J)** Histograms depicting the distribution of reporter ion intensities for each normalized reporter ion. **H)** Box and whiskers plot depicting the distribution of data for each reporter. Each reporter has been normalized to center the data around zero by subtracting the median. **K)** Boxplot representing reporter ion intensities across all samples after log transformation and normalization by subtracting the median. No major shifts in mean intensity occurs between reporters.

# Supplementary figure 6




**Supplementary figure 6.** Venn diagram showing the differential expression of proteins when analyzed with the EOT relapse (PID 338) case included as a TB case (with relapse), excluded from the analysis (without relapse) and included as an EOT case (relapse as EOT)

**Supplementary table 1:** Clinical and microbiological parameters of pulmonary TB patients (PID) assessed after successful anti-TB therapy (M6) using the Differentially culturable tubercle bacilli (DCTB) assay. Age in years, median (range), diagnosis (Dx), body mass index (BMI), chest x-ray (CXR), time to positivity (TTP), time to negativity (TTN), BALF (BLF), induced sputum (iS) post-bronchoscopy sputum (pbS).

| PID | Age | Sex | HIV | Hba1c | Smoker | Previous TB | Treatment adherence (%) | BMI Dx | BMI M6 | CXR M6 | CXR M18 | Culture Dx TTP (days) | Culture TTN (weeks) | DCTB | Source | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 250 | 43 | M | NEG | 6,1 | Daily | NO | 100 | 16,8 | 17,8 | Improved | Improved | 8 | 8 | | | Cured |
| 253 | 29 | M | NEG | 5,9 | Daily | NO | 100 | 19,2 | 21,3 | Worse | Improved | 11 | 16 | POS | BLF | Cured |
| 267 | 57 | M | NEG | 6 | EX | YES | 100 | 18,2 | 21 | No change | No change | 6 | 16 | POS | BLF/iS | Cured |
| 273 | 32 | M | POS | 5,7 | Daily | NO | 100 | 18,1 | NA | Improved | Improved | 7 | 8 | | | Cured |
| 278 | 36 | M | NEG | 5,4 | NO | YES | 99 | 20,8 | 21,6 | Improved | No change | 7 | 8 | | | Cured |
| 291 | 33 | M | NEG | 5 | Daily | NO | 99 | 16,6 | 19,4 | No change | Improved | 5 | 24 | | | Cured |
| 300 | 40 | M | NEG | 5,3 | Daily | NO | 100 | 20 | 20,2 | Improved | No change | 4 | 2 | | | Cured |
| 306 | 55 | M | NEG | 5 | Daily | NO | 100 | 20,9 | 21,2 | No change | Improved | 5 | 8 | | | Cured |
| 314 | 42 | F | NEG | 5 | Daily | NO | 100 | 22,6 | 24,2 | Improved | No change | 16 | 8 | | | Cured |
| 315 | 26 | F | NEG | 5,6 | Daily | NO | 100 | 17 | 17,4 | Improved | No change | 4 | 24 | | | Cured |
| 318 | 31 | F | NEG | 5,9 | NO | NO | 100 | 21,5 | NA | Improved | Improved | 16 | 16 | | | Cured |
| 332 | 58 | M | NEG | 6,1 | Daily | NO | 100 | 21 | 23,9 | Improved | NA | 5 | 8 | | | Cured |
| 334 | 39 | M | NEG | 5,9 | EX | YES | 100 | 16,1 | 16,3 | Improved | Worse | 9 | 4 | | | Cured |
| 335 | 49 | M | NEG | 5,9 | Daily | YES | 100 | 15,4 | 17,2 | Improved | No change | 5 | 8 | | | Cured |
| 338 | 48 | M | NEG | 5,3 | Daily | YES | 99 | 16 | 16,5 | No change | No change | NA | 8 | POS | BLF/iS/pbS | Relapse |
| 351 | 22 | F | NEG | 5,5 | Daily | NO | 100 | 17,9 | 20,9 | Improved | Improved | 5 | 8 | | | Cured |
| 354 | 62 | F | NEG | 6 | Daily | YES | 100 | 27,4 | 29,8 | Improved | Improved | 6 | NA | | | Cured |
| 357 | 51 | M | NEG | 5,2 | Daily | YES | 100 | 18,7 | 18,4 | Improved | No change | NA | NA | | | Cured |
| 359 | 48 | F | NEG | 5,5 | Daily | YES | 100 | 22,1 | 23 | Improved | No change | 16 | 2 | | | Cured |
| 365 | 65 | M | NEG | 5,2 | < daily | YES | 100 | 20,6 | 20,5 | Improved | No change | NA | NA | | | Cured |
| 370 | 26 | F | NEG | 5,2 | < daily | NO | 100 | 16,5 | 18,6 | Improved | No change | 4 | 8 | POS | iS | Cured |
| 375 | 54 | M | NEG | 6,3 | Daily | NO | 100 | 15,2 | 17,8 | Improved | Improved | 10 | 8 | | | Cured |

**Supplementary table 1:** Clinical and microbiological parameters of pulmonary TB patients (PID) assessed after successful anti-TB therapy (M6) using the Differentially culturable tubercle bacilli (DCTB) assay. Age in years, median (range), diagnosis (Dx), body mass index (BMI), chest x-ray (CXR), time to positivity (TTP), time to negativity (TTN), BALF (BLF), induced sputum (iS) post-bronchoscopy sputum (pbS). (*continued*)

| PID | Age | Sex | HIV | Hba1c | Smoker | Previous TB | Treatment adherence (%) | BMI Dx | BMI M6 | CXR M6 | CXR M18 | Culture Dx TTP (days) | Culture TTN (weeks) | DCTB | Source | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 382 | 43 | F | NEG | NA | Daily | NO | 99 | 17,2 | 18,7 | Improved | Improved | 4 | 16 | | | Cured |
| 392 | 21 | M | NEG | NA | Daily | NO | 100 | 17,5 | 17,9 | Improved | No change | 11 | 8 | | | Cured |
| 393 | 39 | M | NEG | NA | Daily | NO | 96 | 16,3 | 17,9 | Improved | Improved | NA | 16 | | | Cured |
| 394 | 44 | F | NEG | NA | Daily | YES | 100 | 25,6 | 26,8 | Improved | Improved | NA | NA | | | Cured |
| 396 | 52 | M | NEG | NA | Daily | NO | 100 | 20,1 | 22,5 | No change | Improved | 5 | 8 | | | Cured |
| 397 | 47 | M | NEG | NA | Daily | YES | 96 | 20,3 | 19,6 | No change | Improved | NA | NA | | | Cured |
| 404 | 49 | F | NEG | NA | NO | YES | 99 | 16,6 | 15,6 | No change | No change | 8 | 24 | POS | iS | Relapse |
| 413 | 29 | M | NEG | NA | Daily | NO | 100 | 18,5 | NA | Improved | Improved | 8 | 4 | | | Cured |
| 426 | 33 | F | NEG | NA | NO | NO | 100 | 22 | NA | Improved | Improved | 9 | 8 | | | Cured |
| 425 | 30 | M | NEG | NA | Daily | NO | 100 | 19,2 | NA | Improved | Improved | 13 | 4 | | | Cured |
| 422 | 57 | M | NEG | NA | Daily | NO | 100 | 23,3 | NA | Improved | Improved | 16 | 2 | | | Cured |
| 427 | 51 | M | NEG | NA | Daily | NO | 99 | 17,2 | NA | Improved | Improved | 9 | 4 | | | Cured |
| 428 | 32 | M | NEG | NA | Daily | NO | 100 | 15 | NA | Improved | Improved | 7 | 8 | | | Cured |
| 429 | 52 | M | NEG | NA | Daily | YES | 100 | 15,9 | NA | Improved | Improved | 6 | 8 | | | Cured |
| 430 | 45 | M | NEG | NA | Daily | NO | 100 | 18,6 | NA | Improved | Improved | 13 | 8 | | | Cured |
| 431 | 23 | F | NEG | NA | < daily | NO | 100 | 20,4 | NA | Improved | Improved | 5 | 16 | | | Cured |
| 390 | 54 | F | NEG | NA | NO | NO | 99 | 21,4 | 19,3 | Improved | Improved | 11 | 8 | | | Cured |
| 437 | 23 | F | NEG | NA | Daily | NO | 100 | 20,8 | NA | Improved | Improved | 5 | 4 | | | Cured |
| 433 | 31 | M | NEG | NA | Daily | NO | 100 | 26,1 | NA | Improved | Improved | 8 | 2 | | | Cured |
| **Total 41** | **43 (21-65)** | **27 Male 66%** | **40 HIV Negative 97.5%** | | **34 Smoker 82.93%** | **13 Previous TB 31.7%** | **>99%** | **19 (15-27.4)** | | **33 Improved 80%** | **26 Improved 63%** | **7 (4-16)** | **8 (2-24)** | **5 DCTB Positive 12.2%** | | **2 relapse 4.88%** |

# 7

# ProVision: A web based platform for rapid analysis of proteomics data processed by MaxQuant.

James Luke Gallant[1,2]
Tiaan Heunis[1,3]
Samantha Leigh Sampson[1]*
Wilbert Bitter[2,4*]

[1]DST/NRF Centre of Excellence in Biomedical TB research, SA MRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Department of Biomedical Science, Faculty of Medicine and Health Science, Stellenbosch University, Tygerberg, 7505, Cape Town, South Africa
[2]Section Molecular Microbiology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam 1081 HZ, Amsterdam, The Netherlands
[3]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom,
[4]Medical Microbiology and Infection Control, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam 1081 HZ, The Netherlands

Bioinformatics

DOI: 10.1093/bioinformatics/btaa620

# ABSTRACT

## Summary

Proteomics is a powerful tool for protein expression analysis and is becoming more readily available to researchers through core facilities or specialised collaborations. However, one major bottleneck for routine implementation and accessibility of this technology to the wider scientific community is the complexity of data analysis. To this end, we have created ProVision, a free open-source web-based analytics platform that allows users to analyse data from two common proteomics relative quantification workflows, namely label-free and tandem mass tag-based experiments. Furthermore, ProVision allows the freedom to interface with the data analysis pipeline while maintaining a user-friendly environment and providing default parameters for fast statistical and exploratory data analysis. Finally, multiple customisable quality control, differential expression plots as well as enrichments and protein-protein interaction prediction can be generated online in one platform.

## Availability and implementation

Quick start and step-by-step tutorials as well as tutorial data is fully incorporated in the web application. This application is available online at https://provision.shinyapps.io/provision/ for free use. The source code is available at https://github.com/JamesGallant/ProVision under the GPL version 3.0 license.

## INTRODUCTION

Mass spectrometry-based shotgun proteomics is a powerful tool that allows researchers a means to investigate the proteome of an organism in an unbiased manner. However, the data analysis associated with proteomics often has a steep learning curve and thus presents a barrier for first-time users.

To address this, tools such as Perseus (1,2), LFQ-analyst (3) and various R packages (4–7) have been created. Perseus is currently a widely used companion tool for analysing data from the popular MaxQuant proteomics analysis platform (8). Perseus provides a wealth of functionality to interface with various label-free and label-based proteomics experiments. However, the proteomics data analysis pipeline can be daunting to newcomers and requires a significant time investment as it is not immediately evident which steps are required. Alternative R-based tools such as Proteus, LFQ-analyst, and MSstats (4) provide the usability of powerful open source tools from the R-language with a focus on allowing users to analyse data in an automated approach. However, either knowledge of the R-language is required or only data that incorporates the maxLFQ (9) algorithm is supported. Furthermore, automated data analysis provides an attractive option when data is routinely analysed, but may not be beneficial on a per-use basis where altering key parameters can result in in various statistical outcomes.



Here we have created ProVision, a web-based and user-friendly proteomics data analysis platform for downstream analysis of MaxQuant output. The platform currently supports label-free data with and without the maxLFQ algorithm as well as tandem mass tag (TMT) data. Importantly, ProVision has been created to complement the reactive nature of the R-shiny web framework. Therefore, users can interact with important filtering and statistical parameters and view the effects in real time as the changes propagate through the platform. Default parameters are provided to guide unfamiliar users through the analytical steps, thereby addressing a potential learning curve for new users. In addition, ProVision aims to consolidate the proteomics data analysis workflow in one platform by providing built-in functionality to perform hypothesis tests, pathway and gene ontology enrichment using Webgestalt as well as protein-protein interactions using STRING (10–12). By use of this platform, the time spent learning and performing proteomics data analysis is reduced and biologically relevant conclusions can be reached within one platform. ProVision thus provides an exploratory data analysis platform where users can gain insight into their data and tweak parameters when needed, reach biologically relevant conclusions through hypothesis testing and enrichment analysis as well as create custom figures with a high degree of control within the browser. This has the potential to dramatically decrease

the turnaround time for proteomics experiments, resulting in faster conclusions and accelerated discovery.

## RESULTS

ProVision was created with the use of the R-shiny web platform and styled with shinydashboard, shinyjs and shinywidgets as well as custom HTML/CSS and JavaScript to create a user-friendly experience.

The platform requires the proteingroups.txt file from MaxQuant and processes data by extracting relevant columns based on the experiment type, while removing identifications flagged as contaminants. From this point, the data is fully reactive and each filtering parameter has a default value that can be changed based on user preference. With this feature it is possible to experiment with critical parameters that will effect the outcome, while simultaneously retaining statistical rigour. Thereby providing end-users with a dynamic data analysis pipeline for specific use cases. Multiple quality control plots such as Q-Q plots (Fig. 1A), Histograms (Fig. 1B), Scatterplots (Fig. 1C), correlation heat maps (Fig. 1D), and principal component analysis (Fig. 1E) can be created.

Key figures that can be created include volcano plots (Fig. 1F) and heat maps (Fig. 1G) from statistically significant data. Analysis of the protein lists can be further extended by performing gene set enrichments and over-representation analysis using Webgestalt (10,11). Furthermore, both known and predicted protein-protein interactions can be done within the application using the STRING database (12). These analysis are done directly on the platform and automatically integrates with all upstream data analysis. The plots are created by ggplot2 and thus have multiple customisable options. Furthermore, all plots can be exported to major file formats and can be exported to PDF for vector graphics and downstream processing. The statistical analysis is done using the Limma (13) package and is fully reactive as well. This allows for any statistical changes made to propagate to the volcano plots, heatmaps, Webgestalt enrichments and STRING networks in real-time and update the display. The differential expression as well as the analysed data can be downloaded in either Excel or text format, with a choice of various delimiters, for downstream analysis. Finally, no uploaded data is stored within the server thus creating a safe environment with the caveat that progress can be lost if the browser window is closed. Both quick start and full tutorials are available online and embedded within the application for users to access. ProVision is under continuous development with source code available to advanced users who would like to contribute or request specific features.

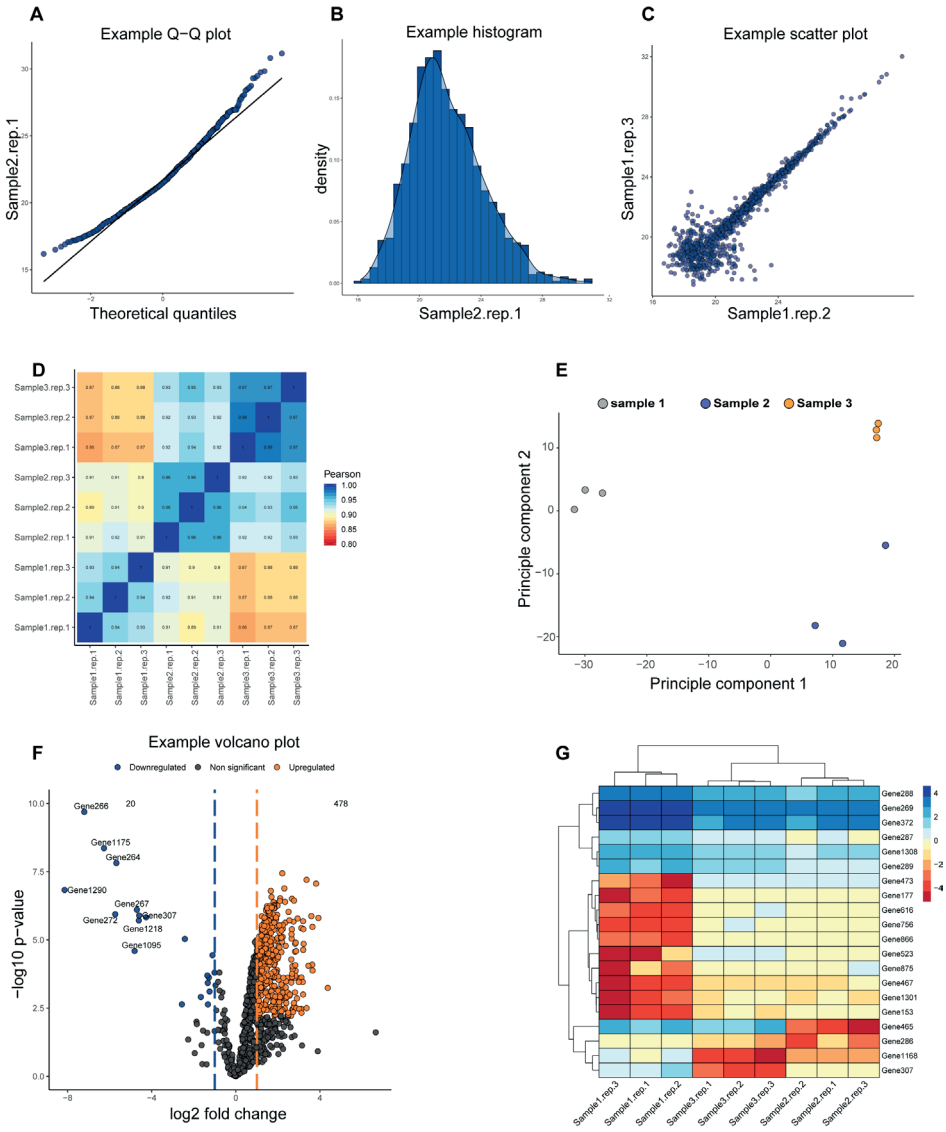



**Figure 1:** Representation of the various graphs that can be created by Provision using the tutorial data. Quality control plots include **A)** Q-Q plots, **B)** histograms, **C)** scatterplots, **D)** correlation heat maps and **E)** principle component analysis. Main figures include **F)** volcano plots and **G)** heat maps. Multiple parameters of every plot can be customised to user preference.

## CONCLUSIONS AND OUTLOOK

ProVision is an open source web application designed for ease of use and accessibility to newcomers for proteomics data analysis. ProVision aims to assist researchers to reach accurate conclusions based on their unique experimental designs, while providing high-quality customisable graphs and statistics in an intuitive environment. In addition, users can revisit their analysis and change parameters to gain the optimal output if necessary as well identify differentially regulated proteins, pathways and networks. This platform is deployed at https://provision.shinyapps.io/provision/ for general use and a development version is available at https://github.com/JamesGallant/ProVision for advanced users who would like to contribute to its development.

## FUNDING

This work was supported by the NRF-VU Desmond Tutu Doctoral training program for financial assistance to JG. The authors acknowledge the SA MRC Centre for TB Research and DST/NRF Centre of Excellence for Biomedical Tuberculosis Research for financial support for this work. SLS is funded by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa, award number UID 86539.

## AUTHOR CONTRIBUTIONS

**JG:** Conceptualisation, Software, methodology, project administration, writing

**TH:** Software, validation, methodology, writing

**WB:** Supervision, writing

**SLS:** Supervision, writing

## REFERENCES


1. Tyanova S, Cox J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. In: Methods in molecular biology (Clifton, NJ) [Internet]. 2018 [cited 2019 Oct 13]. p. 133–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29344888
2. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods [Internet]. 2016 Sep 27 [cited 2018 Mar 11];13(9):731–40. Available from: http://www.nature.com/articles/nmeth.3901
3. Shah AD, Goode RJA, Huang C, Powell DR, Schittenhelm RB. LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant. J Proteome Res [Internet]. 2020 Jan 3 [cited 2020 Jan 20];19(1):204–11. Available from: https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00496
4. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014 Sep 1;30(17):2524–6.
5. Gatto L, Lilley KS. Msnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics. 2012 Jan;28(2):288–9.
6. Gatto L, Breckels LM, Naake T, Gibb S. Visualization of proteomics data using R and Bioconductor [Internet]. Wiley-VCH Verlag; Apr, 2015 p. 1375–89. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pmic.201400392
7. Gierlinski M, Gastaldello F, Cole C, Barton GJ. Proteus: an R package for downstream analysis of MaxQuant output. bioRxiv [Internet]. 2018 [cited 2020 Jan 21];416511. Available from: http://dx.doi.org/10.1101/416511%0Ahttps://www.biorxiv.org/content/10.1101/416511v2
8. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
9. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics [Internet]. 2014 Sep [cited 2019 Oct 13];13(9):2513–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24942700
10. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res [Internet]. 2017 Jul 3 [cited 2019 Mar 7];45(W1):W130–7. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx356
11. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res [Internet]. 2019 [cited 2020 May 26];47:199–205. Available from: https://academic.oup.com/nar/article-abstract/47/W1/W199/5494758
12. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res [Internet]. 2018 [cited 2020 May 26];47:607–13. Available from: https://string-db.org/.
13. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Apr 20 [cited 2018 Mar 15];43(7):e47–e47. Available from: http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for

# 7

# Addendum

This addendum acts as an extended report on the software presented in the manuscript titled: ProVision: A web based platform for rapid analysis of proteomics data processed by MaxQuant. Parts of this addendum will contain some of the tutorial explanations as well.

## BACKGROUND

The cloud-based application, ProVision, was written in the R-language using the Shiny web frame work to create a data analysis platform geared towards users that are unfamiliar in R as well as proteomics. This is achieved by providing a front end to multiple R functions and packages in an intuitive manner which further allows users to quickly run through an analysis while retaining the ability to tweak settings as desired. The Shiny web frame work allows reactive functionality which forms the central functionality of ProVision, where data created upstream can propagate downstream and influence both the statistics and graphs created, allowing flexibility in analysis while still being limited in order to maintain statistical rigour. It is also possible to use the differential quantification data generated in the statistical tests and perform third party analysis such as Webgestalt (1) and STRING (2) enrichments by querying their respective public application programming interfaces (API). Finally, a tutorial system was developed and incorporated as well which aims to assist potential newcomers to the various steps associated with proteomics data analysis.



## MAIN WORKFLOW

ProVision uses a linear data analysis flow which is represented by tabs and constitutes the main workflow. Each tab is subsequently sub-divided into smaller units which govern the central functionality of ProVision. The main workflow consists of four sections, namely data handling, quality metrics, statistics and Figure construction (Figure 1). Each of these will be covered in detail in the following sections.

### Data handling

#### Uploading

ProVision currently accepts input from proteomics data searched with MaxQuant software (3), thus the proteingroups.txt file. This application is further compatible with both label-free and tandem mass tagged proteomics experiments. The application will also restructure all the default values depending on which experiment was done by the user.

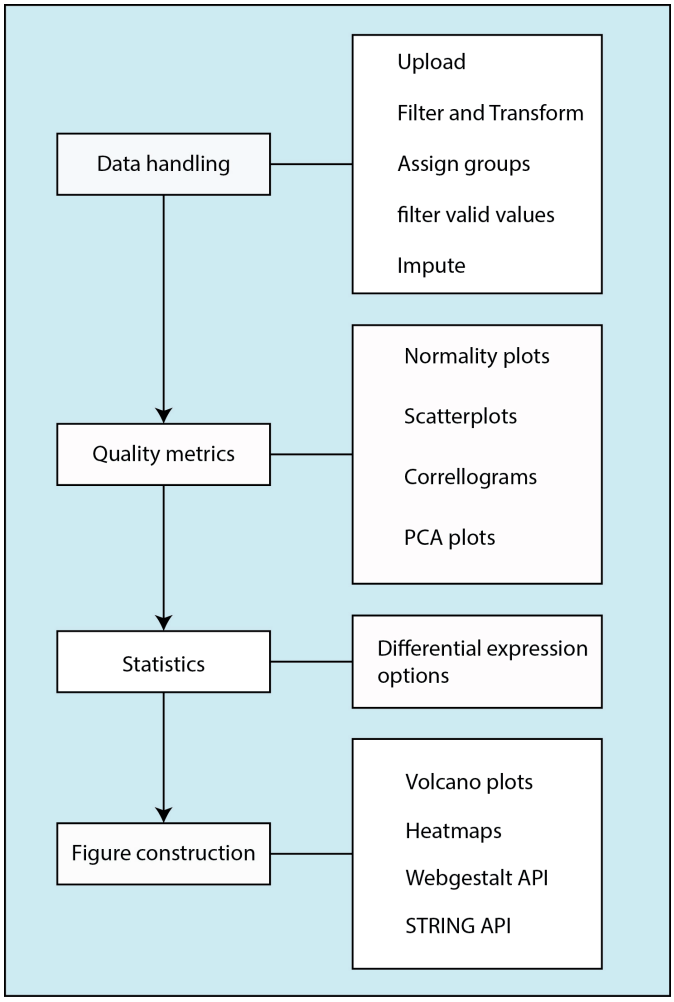



**Figure 1:** The complete ProVision workflow detailing the various components and steps in the application.

## Filter and transform

In this section the raw intensity data is transformed to remove flagged false positives and potential false positives based on peptide counts. MaxQuant provides convenient filtering methods that are readily available for automatic processing, these include identifications that are only present due to site modifications; reverse database hits and potential contaminants. Proteins identified by their modifications alone, such as oxidation on methionine or N-terminal acetylation, have no peptides that indicate that this protein was specifically identified, therefore they are typically removed. The potential contaminants and reverse database hits are candidates that have been identified as common contaminating proteins, such as albumin or keratin, and proteins that have been identified in the reverse amino acid sequence. In both of these cases the

proteins flagged as such are removed to reduce the noise within the dataset as well as the number of proteins contributing to the statistical test which is performed later in the pipeline. Another metric to consider as potential false positive is the number of unique peptides attributed to an individual protein. When performing peptide spectrum matching, proteins are either assigned unique peptides or razor peptides. The unique peptides are lengths of amino acids that can be matched to an specific protein and thus provides a level of confidence towards annotating the presence and abundance of a specific protein. Razor peptides are named for the principles of Occam's razor or the law of parsimony, where protein hits may be reported as the minimum set that accounts for all observable peptides. The unique peptides thus provide the most compelling evidence for the presence of a protein within the sample set. To add to this confidence, the protein groups with less than 2 unique peptides are filtered away as well. This parameter is changeable within ProVision, and increasing this number will yield more confident hits but at the cost of abundance and increases the risk for false negatives. By introducing these filtering steps the initial dataset has an increased confidence and reduces the number of hypothesis tests which will be discussed later. By reducing the number of tests there is a lower chance for introducing false positives and the multiple hypothesis correction algorithm will have less tests to constrain resulting in a lower number of false negatives.



As part of the initial filtering it is also possible to transform data, which is standard and centre data around the mean intensity. In ProVision, parametric hypothesis testing is used which has a number of constraints. While these constrains are (at times) soft constrains, generally the test performs the best when they are met. The most important is that data should be normally distributed. Typically, biological data follows a log-normal distribution which is not favoured by parametric statistical tests, as is done within ProVision, and would not contain additional distribution (i.e. not binomial). Thus by log transforming the intensities, a normal distribution is achieved and the data is suitable for parametric testing. Another option is to centre the median around zero which reduces variation between groups. This may be useful to reduce variation on the mass spectrometry itself and increases confidence in hypothesis tests without significantly altering the intragroup variation. The function used to centre the median is shown below:

```
center_median = function(x) {
  kolom_name <- as.data.frame(table(unlist(names(x))))
  kolom_name <- as.character(kolom_name $Var1)

# Iterates the column name character list
  x[, kolom_name] = lapply(kolom_name,
                           function(i){
                             LOG2 = x[[i]]
                             LOG2[!is.finite(LOG2)] = NA
                             gMedian = median(LOG2, na.rm = TRUE)
                             LOG2 - gMedian
                           })
```

**Assign groups**

This section is simply a user input section to define groups, which ProVision handles internally for a multitude of functions. The user provides the groups corresponding to each protein ID and also has an option to provide axis names as well. These groups are bound to the column and used to determine characteristics of an experiment automatically such as the number of groups present and which unique experiments map to which group.

**Filter valid values**

This step is used to apply experiment specific logic and filter the data further and thus create an optimally reduced dataset for controlling the false discovery rate. Filtering for valid values simply means filtering for missing values, during log transformation these will be labelled as NA and thus register as non-valid. These values can occur due to two main factors, absent across replicates or absent due to experimental conditions.

Addressing the first case, it is completely possible that the spectra contributing to a specific protein are missed across different mass spectrometry runs or different replicates of a condition. This does not mean that this protein is missing, it simply means that the mass spectrometer did not observe it in a specific run. To address this, it is possible to filter out proteins that were seen at least twice in a total of three replicate runs. Thus we are confident that this protein exists yet allow for some lenience with the knowledge of how the mass spectrometer operates. Changing the number to three (if you have three replicates) would only consider proteins that were observed in each replicate of a sample and thus be the strict case.

In the second case, it can be conceived that certain proteins will not be expressed in response to a condition, as per example addition of a drug in a cell line. This missing

data is thus important to keep in the calculations as it likely contains the proteins activated in response to the drug. To do this we choose in at least one group where groups were defined in the assigned groups section. By choosing the in each group option the minimum value filtering parameter will be applied to each group defined in the assigned groups section. To achieve the strictest possible filtering, assign the minimum values equal to your number of replicates and choose in each group. While it might be tempting to keep data as strict as possible, using this filtering parameter discards a large amount of usable data which could contribute to your phenotype. However, if you do decide to filter at the most strict level there will be no residual missing values and you are done with the section and imputation can be skipped.

### Impute missing values

Missing values can be detrimental to statistical test, while it can be handled internally within an hypothesis test, it typically does not perform as well when the data is present. Depending on the type of filtering done there can be missing values present within the dataset prior to the statistical test and these can be filled in using imputation. Imputation refers to the process of including missing data within a given data set by extrapolating information from the data and making an "educated guess" of sorts. While many algorithms aim to do so, i.e. K-nearest neighbours, random forests or linear regression. These methods often assume that data is missing at random within the dataset, thus the missing data is imputed based on the current shapes and distributions of the data and not of the experimental setup that created the data. The simplest imputation method that follows this rule is by imputing missing values with the mean of the distribution. It should be immediately apparent that such an approach is not recommended for proteomics or any expression data as these values are not randomly missing and inclusion of missing values as a mean of the total distribution would skew data to the centre and no differences would be detected proteins who are perhaps low and mid abundant within a group if enough values are missing.

Missing values in a proteomics dataset can arise from two main methods, either the mass spec did not observe the spectra associated with a specific protein or the protein was simply not expressed. In both cases it is highly likely that the missing value falls within the low range of the distribution and not the mid-range. Thus one of the popular approaches is to extract the lower quantiles of the normal distribution in which the specific protein falls (Figure 2). The lower quantile represents the low abundant proteins. Regardless if it is not expressed or too low to be detected, the hypothesis is that the value would fall within the lowest quantile of the total distribution. Therefore by choosing a value at random in the lower quantile, a low expression protein is simulated for every missing value.

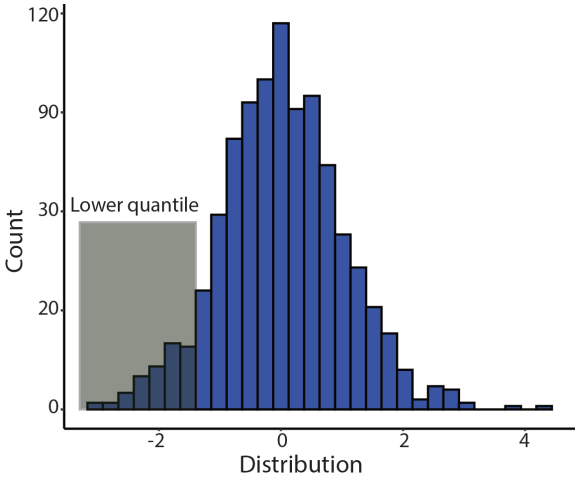



**Figure 2:** The lower quantile of a normal distribution. This quantile is used to infer missing values associated with low protein expression and other missing values.

There are limitations, such as not taking into account replicate values, however this approach performs well in combination with the filtering parameters described above.

After filtering and imputation the data is considered a high confidence data set and can be inspected in the QC tab or can be used for statistical tests directly. Any changes made within these critical steps will propagate throughout the analysis, therefore users can return to this section and change parameters as required.

## Quality metrics

This section is used to evaluate metrics surrounding the proteomics data and checks prior to hypothesis testing can be done here. We focussed on metrics that has an influence on both the statistical test as well as the experimental setup. The available metrics are quantile of quantile plots, histograms, scatter plots, correlation heatmaps and principle component analysis. The following sections will provide an explanation on the interpretation of each plot.

## Quantile of quantile plots and histograms

Quantile of quantile (Q-Q) plots and histograms can be used to visualise a normal distribution. Ideally data should be normally distributed for imputation as well as statistical testing. Q-Q plots take the given data, in this case intensity data after processing and compares it to a known distribution using a scatter plot. The theoretical quantile represents a known normal distribution compares the quantiles of this distribution to the quantiles of each sample (Figure 3). The known normal distribution is calculated internally in the application. If a perfect straight line is formed then data is considered100% normally distributed. Typically tailing can be seen, either one or two tails and this is normal for proteomics data. Histograms function in a similar manner and plots the distribution curve. Tailing is not as easily spotted in histograms but a clear indication of normality can be observed using this method. If distributions are all you need then this should

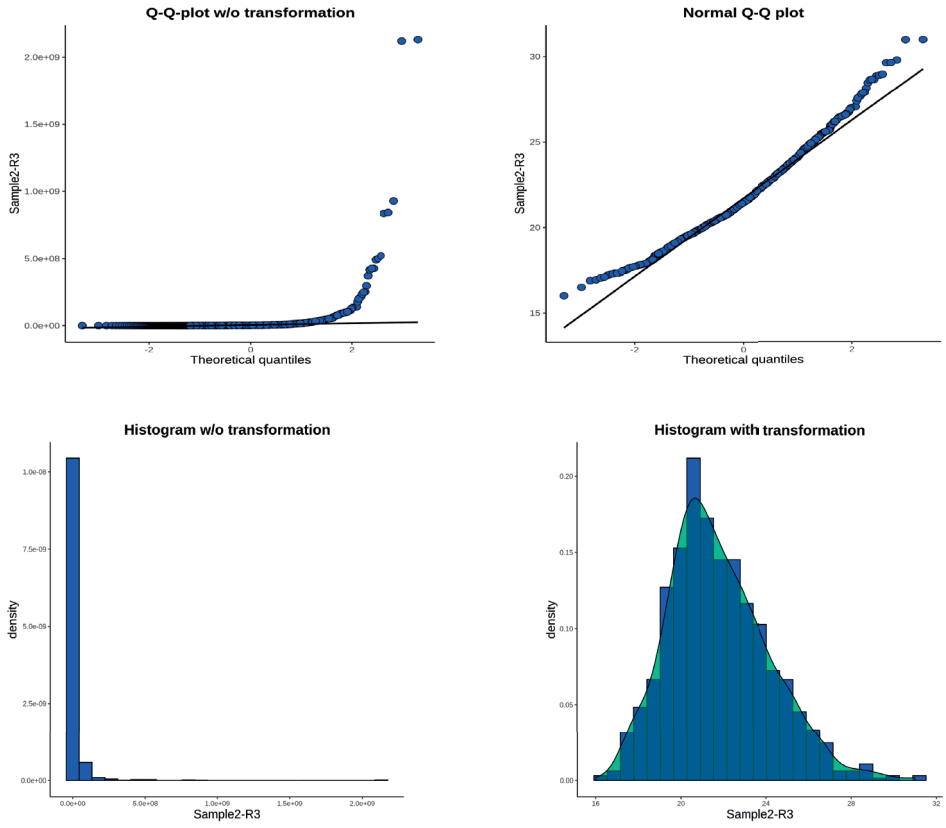



**Figure 3:** Example of the quantile of quantile plots as well as the associated histogram to demonstrate the effect of log transformations on the data distribution. Quantile of Quantile plots are represented with a theoretical normal distribution.

```
NormalityPlot <- reactive({
    # Defensive check
    if (is.null(input$user_file)) {
      return(NULL)
    }

    processed_data <- processed_data()

    processed_data$GeneNames <- NULL

    anno_data <- anno_data()
    colnames(processed_data) <- anno_data$axisLabels
    ylabname <- colnames(processed_data[Counter$normcounter])
    index <- Counter$normcounter
    qqplotList <- list()
    histPlotList <- list()

    #init plot cylce on render to save computing time
    if (input$normRender > 0) {
      #run for selected plots only
      if (input$normPlotChoice == "qqPlot") {
        for (i in 1:ncol(processed_data)) {
          dat <- data.frame(qqnorm(processed_data[,i], plot.it = FALSE))
          axisname <- colnames(processed_data[i])
          p <- ggplot(dat, aes(x, y)) +
            geom_point(pch = 21,
                       alpha = input$qqPlotAlphaChannel,
                       size = input$qqPointSize,
                       colour = input$normPlotCol,
                       fill = input$normPlotFill) +
            geom_qq_line(aes(sample = y),
                         lty = input$qqLinesType,
                         lwd = input$qqLineWidth,
                         colour = input$qqLineCol) +
            ylab(axisname) +
            xlab("Theoretical quantiles") +
            ggtitle(input$qqPlotTitle) +
            theme_classic(base_size = 14) +
            theme(plot.title = element_text(hjust = input$normTitlePos,
                                            face = input$normTitleFace,
                                            size = input$normTitleSize),
                  axis.title.x = element_text(size = input$normXsize),
                  axis.title.y = element_text(size = input$normYsize))

          qqplotList[[i]] = p
        }

        return(qqplotList)
```

### Scatter plot

Drawing of the scatterplots:

```
scatterplot <- reactive({
    if (is.null(input$user_file)) {
      return(NULL)
    }

    processed_data <- processed_data()
    anno_data <- anno_data()

    anno_data$axisLabels <- sapply(anno_data$axisLabels, function(data){
      data <- str_replace(pattern = "-
", replacement = ".", string = data)
      return(data)
    })

    colnames(processed_data) <- as.character(anno_data$axisLabels)
    index <- Counter$scatcounter
    plot_list <- list()
    plot.col <- input$scat_point_col
    for(i in unique(anno_data$annotation)){
      COLS=anno_data$axisLabels[anno_data$annotation ==i]
      plot_combinations <- combn(COLS,
                                 2,
                                 simplify = FALSE)

      for (a in 1:length(plot_combinations)) {
        p = ggplot(processed_data,
                   aes_string(x = plot_combinations[[a]][1],
                              y = plot_combinations[[a]][2])) +
          geom_point(pch = as.integer(input$scatCharcters),
                     colour = input$scatPointBorder,
                     size = input$scatPointSize,
                     fill = plot.col,
                     alpha = input$scatPlotAlphaChannel) +
          ggtitle(input$scatTitle) +
          theme_classic(base_size = 14) +
          theme(plot.title = element_text(hjust = input$scatTitlePos,
                                          face = input$scatTitleFace,
                                          size = input$scatTitleSize),
                axis.title.x = element_text(size = input$scatXsize),
                axis.title.y = element_text(size = input$scatYsize))
        out_name <- paste(i,a,sep = "_")
        plot_list[[out_name]] = p
      }
    }

    if (input$scatRender == 0) {
      return(NULL)
    } else {
      dispPlot <- plot_list
```

Scatter plots are used to visualise comparisons of one sample to another. Ideally there should be a correlation between replicates and lower correlation within replicates, scatter plots are used to visualise this correlation (Figure 4). If there is a linear increase on the x and y axis, then samples are positively correlated as show in the Figure. Low correlation can be indicative of an outlier sample that should rather be left out from further calculation as it will highly skew the standard deviation. These are the more likely sce-narios which are found within scatterplots. It is also possible to have a negative correlation, where there is a linear decrease of x and y. Scatterplots provide a good metric to evaluate a sample compared to another sample by visualising the proteins comparatively in each strain. Intragroup comparisons are not as informative using scatter plots and thus were restricted in ProVision, correlation heatmaps provide a better metric for inter- and intragroup comparisons.

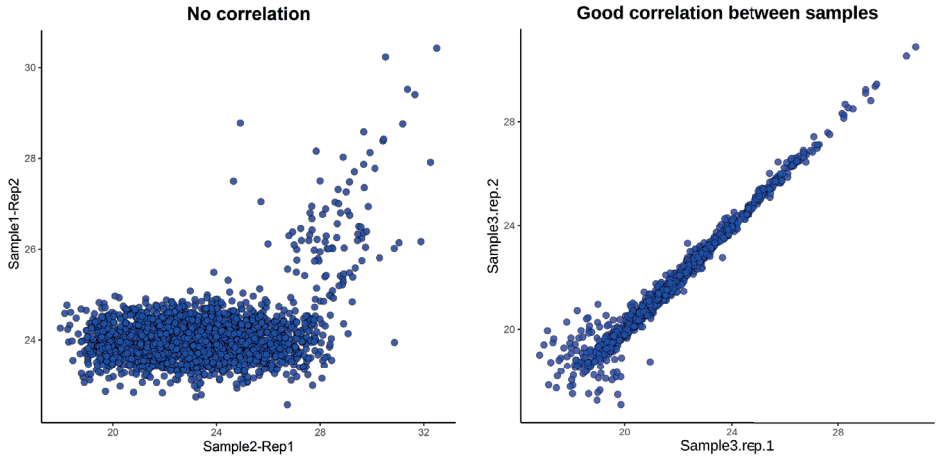



**Figure 4:** Scatter plots showing the correlation of proteins between samples. Straight lines represent correlated samples.

## Correlation heatmaps

Correlation heatmaps are used to visualise the extent to which two variables change together, i.e. the correlation coefficient. The correlation coefficient is simply a numerical measure of correlation, in ProVision this correlation coefficient can be calculated using either Pearson or Spearman methods. This value is tied to the scatter plots as displayed in the previous section and essentially measures the extent to which two variables change together. Correlation coefficients describe both the strength and direction of a given relationship. If a positive perfect correlation is achieved this number would be 1 or -1 for a perfect negative correlation and the associated scatter plot would present a perfect straight line. Instead of drawing scatter plots for each comparison, a correlation heatmap can visualise these relationships using a scaled values from -1

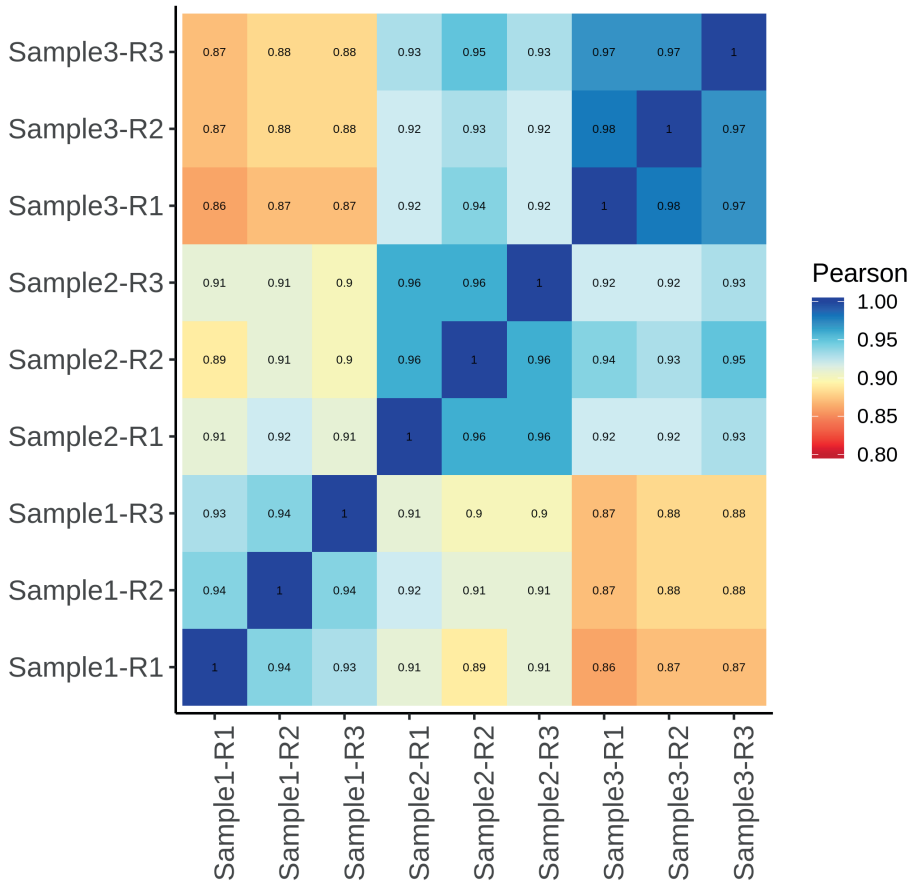



**Figure 5:** Correlation heatmap plotting the Pearson or Spearman correlation. This heatmap reduces the dimensionality of multiple scatter plots in order to gain higher order information on the inter-sample relationships.

to 1. In the context of proteomics, this relationship would typically be positive and in ProVision we have constricted this value to range from 0 to 1.

Pearson correlation coefficients measure the relationship between two continuous variables where the change in one variable leads to proportion-al change in another. Spearman correlation measures monotonic relationships between continuous variables. Thus Spearman is able to measure change where variables can change together but not at a constant rate and is based on ranking rather than raw values. In proteomics we hope to see high correlation within groups and low correlation between groups.

This is a strong indication of a good experiment that will likely give statistical significant difference when subjected to hypothesis test.

Drawing correlation heat maps:

```
correllation_heatmap <- reactive({

    if (input$corrRender == 0) {
      return(NULL)
    } else {
      cordata <- corrProcessing()
      p = ggplot(cordata, aes(x=Var1, y=Var2, fill=value)) +
        geom_tile() +
        scale_fill_gradientn(colours = brewer.pal(input$corrColChoice,
                                                  n = input$corrColourScale),
                             name = "Pearson",
                             limits = c(input$corrSlider, 1)) +
        xlab(NULL) + ylab(NULL) + ggtitle(input$corrPlotTitle) +
        theme_classic(base_size = 14) +
        theme(plot.title = element_text(hjust = input$corrTitlePos,
                                        face = input$corrTitleFace,
                                        size = input$corrTitleSize),
              axis.text.x = element_text(angle=90, hjust = 1, size = input$corrXSize),
              axis.text.y = element_text(size = input$corrYSize))

      if (input$corrValDisp == TRUE) {
        txtsize <- par('din')[2] / 2
        p =  p + geom_text(label=cordata$value, size=txtsize * 0.8, color="grey9")

        return(p)
      } else {
        return(p)
      }})
```

### Principal component analysis

Principal component analysis is a test to determine the variation between principle components. These components represent the variation within a given dataset and is similar to correlation but uses stronger metrics by calculating variation directly. The first component represents the majority of variation and the second represents the second most variation and so on. For our purposes the first two components are of importance. If we consider a differential expression experiment, the greatest amount of variation or the first principle component should occur between groups. There is always experimental variation, which often represents the second most variation in an

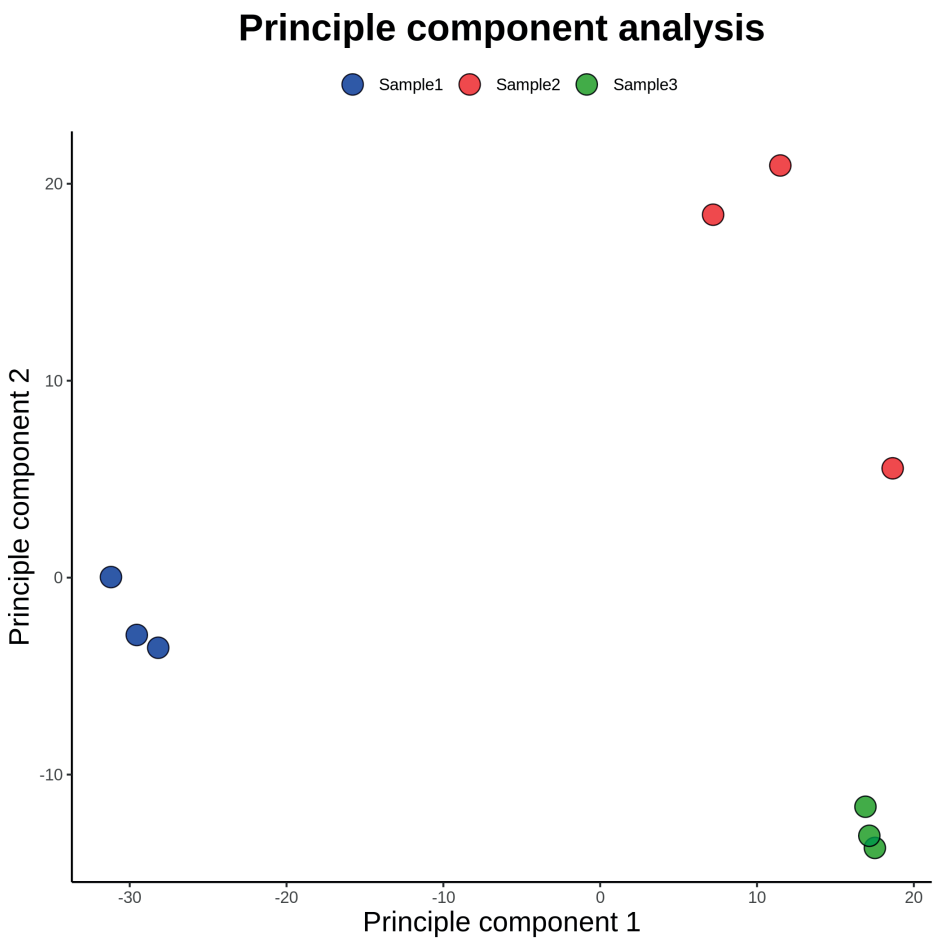



**Figure 6:** Example of a principal component analysis that further reduces the dimensionality of data to represent differences in variation. The first component represents the largest variation and the second component represents the second largest variation.

experiment and thus would cause separation on the second component. In an ideal experimental setup, treatment and control should separate in the first component while replicates of each should separate in the second. Ideally separation between replicates should be minimal and this indicates a good reproducibility of an experiment.

Drawing the PCA plot

```
pca_plots <- reactive({
    pcaData <- pca()

    anno_data <- anno_data()
    cols <- pcaColNames()
    if (is.null(pcaColNames())) {
      cols <- rep("#000000", length(unique(anno_data$annotation)))
    } else {
      cols <- pcaColNames()
    }
    if (input$pcaRender > 0) {
      p <- ggplot(pcaData, aes(pcaData$PC1, pcaData$PC2)) +
        geom_point(pch = as.integer(input$pcaCharcters),
                   size = input$pcaPlotPointSize,
                   colour = "black",
                   alpha = input$pcaAlphaChannel,
                   aes(fill = pcaData$names)) +
        scale_fill_manual(values = cols) +
        theme_classic() +
        ylab("Principle component 2") + xlab("Principle component 1") +
        labs(fill = NULL) +
        ggtitle(input$pcaPlotTitle) +
        theme(legend.position = input$pcaLegendPostition,
              plot.title = element_text(hjust = input$pcaTitlePos,
                                        face = input$pcaTitleFace,
                                        size = input$pcaTitleSize),
              axis.title.x = element_text(size = input$pcaXSize),
              axis.title.y = element_text(size = input$pcaYSize))
      return(p)
    }
  })
```

The PCA analysis provides the option to Z-score the data as the default. This transformation centres the data around zero with a minimal value of -3 standard deviations from the mean and a maximum of +3 standard deviations from the mean. The transformation is done by subtracting each the mean of the total distribution from each observation (protein) and dividing that value by the standard deviation of the distribution. Changing the point size changes the size of the characters, circles by default and can also be changed to other characters. The colour input is generated dynamically based on the input from the assign groups page and can be accessed in a sub menu.

## Statistics and hypothesis testing

The statistical testing parameters can be seen as a research project on its own with many different experiments that can be performed to estimate what the best way would

be to analyse the specific data at hand. Using multiple different methods, we have found Limma (4) to be the best generalized approach for intensity data such as those used in proteomics. The other side is spectral counting which has its own assumptions and we have not tested ProVision for spectral counts nor do we extract spectral counts. As with all other data, the calculations are generalisable and will change if something upstream is changed such as the filtering parameters and so forth. The only thing not fully supported is the removal of samples. The calculation will display the significance counts of each comparison at a time and if more than two groups are present these comparisons can be accessed using the previous and next buttons. In certain cases it may not be beneficial to compare each group to every other group, a time course experiment is a good example of this. Removing unnecessary comparisons is beneficial to the correction algorithms and also the computing time.

### Intensities vs spectral counts

As mentioned, in ProVision Intensities is favoured and the statistical tests used reflect this. For spectral counts, different statistical tests are required. Here we will briefly provide an explanation on why intensities are favoured above spectral counts.

Intensities are based on the precursor intensity or MS1 based measurement by calculating the area under the MS peak or in some cases the maximum intensity of the peak. The sum of these measurements takes into account all the features of the comprised proteins. The precursor ion represents the total intensity of the peptide, but no information on amino acid sequence of the peptide. Thus MS1 is used for quantification and MS2 is used for identification in this model.

Spectral counting is based on MS2 or fragment based measurements where quantification is done by summing the number of identified MS2 spectra and matched against its peptides. The value obtained will depend on the intensity of the protein's precursor peptide ions, as in DDA analyses more abundant features will be sampled more often than lower abundance ones (5).

Spectral counting infers quantification in the same process that identifies the peptides, allows for comparison between samples that are have a high degree of difference and is easy to implement computational. However spectral counting suffers from deficiencies in the method of quantification. The MS1 quantification is much more accurate and has a greater linear dynamic range due to the inherent limitations of spectral counting. With spectral counting the information on the precursor ion is not accounted for which discards potential important characteristics associated with the peak responsible for the peptide as a whole. In spectral counting there is chance for the introduction of

increased internal variation that arises from the fragmentation process as well. Statistically, spectral counting methods typically cannot be used with parametric tests because this type of data rarely follows a normal distribution and suitable transformations need to be tested first. With intensity data, these distributions are typically log normal and the normal distribution can be created by log transformation across all samples. The option for count data is to use non-parametric tests, however statistical power will be lost compared to parametric tests. In proteomics experiments, data is skewed to the most abundant peptides and thus it is important to maximize the statistical power in order to detect small differences within samples. While spectral counting works and can be used to quantify proteins, in ProVision we prefer MS1 based quantification for these reasons.

### Limma and linear models

ProVision uses Limma or linear models for microarray data to calculate significant differences between groups. For in depth information on how Limma works, please refer to their publication (4). Briefly, moderated F-statistics (similar to ANOVA) that combine t-statistics for all comparisons is used as a test for significance in an empirical Bayes approach. Through this process the power of both t-tests as well as F-tests can be utilised and performs exceptionally well with matrix type data. Thus making Limma a powerful choice for proteomics data as well. When comparing Limma to more traditional methods such as ANOVA, the big differences should come from the genes with large variance, and moderate differential expression. This is because the variance estimate is shrunk towards the mean in Limma, while ANOVA uses the sample variance. Since the t or F test uses the variance estimate in the denominator, genes with very small sample variance will be less significant with shrinkage, whereas genes with high variance will be more significant. Finally, powerful correction tools such as Benjamini-Hochberg FDR can be used in conjunction with Limma, which has been shown to work well with proteomics data.

### P-values, Q-values and effect size

The q-value is derived from the p-value by use of a correction algorithm. For proteomics data we have a large amount of observations (i.e. proteins) and thus conducting statistical tests over however many proteins are present within the data set. A typical label free proteomics experiment has around 3000 protein identifications which means an iteration of about 3000 hypothesis tests across multiple groups. If we take into account that a p-value represents a probability value and we set that probability arbitrarily to a value of 0.05 we accept that 95% of a distribution is different and 5% is similar to another distribution. By including more tests this effect is compounded and false positives accumulate. In order to mitigate this effect the p-values are modified by using

correction in order to control the false discovery rate. This correction is known as the q-value and is what we base the call of significance on. The value works similar to p-value and two options are available, either 0.05 for leniency or 0.01 if a stricter cut off is desired. The effect size simply means fold change and is exported as a logarithmic value. The default fold change cut off is set at a log fold change of 1, meaning that there is a one log difference of a protein between two groups. It is wise to include fold change in the cut off metrics when deciding which proteins are significant, especially considering that we work with biological data. Hypothesis tests simply evaluate whether two values and their standard deviations deviate significantly from one another regardless of the magnitude of that difference. However, consider the expression of two proteins. If the absolute concentration of protein A is 1 nM and protein B is 1.5 nM would this be biologically relevant or would a difference of protein A at 1nM and protein B at 1uM represent a stronger biological significance. For this reason we give the option to change the cut-off for fold change significance in a wide range as it is subjective towards a study, the default of one is a good starting point.

### Correction algorithms

To obtain a p-value corrected for performing multiple tests we use correction algorithms. The best performing for the purpose of a proteomics experiment is the Benjamini-Hochberg FDR which provides a good control over the false discovery rate while not being as strict. Benjamini-Hochberg uses a step up method and calculates the correction by ranking each p-value according to its value. Another older, and perhaps more familiar algorithm is the Bonferroni correction. This is strict and increases in stringency as the number of input p-values (proteins) increases because every p-value is only compared to the total number of p-values. With small datasets it is good practice to use Bonferroni as all p-values are treated equally however as we increase the size it is better to control the false discovery rate instead. Thus at high numbers Bonferroni correction would swing the other way and leave false negatives. That being said, for some cases it may be better to have false negative calls than controlling the amount of false positives in the data set. For a more concrete look at p-value correcting within proteomics, a description has been provided by *Diz et al* in their manuscript (6). The Hommel method is very similar to the Benjamini-Hochberg method in that step-up methods are used to control the false discovery rate. The Hommel method is more powerful than Benjamini-Hochberg, however the differences in results is small and the Hommel method is computationally more intensive. The Benjamini-Yekutieli method is the most recent adaptation of the correction methods and is an addition to its predecessor, the Benjamini-Hochberg FDR method. While Benjamini-Hochberg calculates the corrected p-values under certain dependencies (such as positive regression) the Benjamini-Yekutieli method assumes a general dependency. Thus this

method is more generalisable over different data types but it sacrifices power to do so. Probability values corrected with this method would need to be much lower compared to Benjamini-Hochberg FDR to be considered significant yet nowhere close to the strictness of Bonferronni method.

### Performing the statistical tests

To perform the statistical test in ProVision, some key information is required. This includes **1)** knowledge of the groups and their replicates, **2)** performing the test and **3)** extracting the relevant data from the linear models generated.

To identify the groups a matrix is required that displays the comparisons, replicates are mapped to each comparison as defined in the assign groups section. To visualise this simply consider an experiment where there are three groups, namely wild type, mutant and complement The first step is to find all the potential combinations based on the replicates and assigned groups. In our example these combinations are `wild type vs mutant`, `wild type vs complement` and `mutant vs complement`, which has to be calculated without prior knowledge of the number of groups present. This is achieved in the function below by finding the names of the groups form the data defined in the assign groups section and calculating the number of unique annotation names provided by the user. From this information a list of lists can be generated where the second list contains a tuple of each potential comparison.

| Index 1 | Index 2 |
|---|---|
| Wild type | Mutant |
| Wild type | Complement |
| Mutant | Complement |

**Example table of permutations done by ProVision**

These comparisons will be used to generate a contrast matrix used by Limma to calculate differences. This matrix is fed to the front end and displayed to the user and by default selects all comparisons. The inclusion of this choice was to allow sensible choices surrounding the statistical test as more tests equate to more error. Thus if there is a test that is non-sensical it can be removed at this step. The cost of permutation is however set at O(N!) and large comparisons may cause a significant increase in runtime.

```
  statComb <- reactive({
      anno_data <- anno_data()
      comb <- combn(unique(anno_data$annotation), 2, simplify = FALSE)
      contrast <- lapply(comb, function(i) {
        if (input$ComparisonSwitch == 1) {
          compare <- paste(i[1], i[2], sep = "-")
        } else {
          compare <- paste(i[2], i[1], sep = "-")
        }
        return(compare)
      })
    })

 #Send to the front end
  output$statComparisonMat <- renderUI({
      pickerInput(inputId = "hypoTestMat",
                  label = "Choose conditions to compare",
                  choices = unlist(statComb()),
                  multiple = TRUE,
                  selected = unlist(statComb()),
                  options = list(`actions-box` = TRUE,
                                 `selected-text-
format` = "count > 0"), choicesOpt = list(
                                   style = rep(("color: black;"),length(stat
Comb()))))
      )
    })
```

Performing the statistical test is relatively straight forward and can be found in the documentation of Limma. Briefly a matrix is required containing the data to be tested and the column names as well as the group names are provided to helper functions. The statistical test is done in two steps, where the first step fits a model and the second step uses an Empirical Bayes approach to apply the statistical test to the comparisons stated in the previous step. The output is a list of data frames.

```
statsTestedData <- reactive({
    processed_data <- processed_data()
    rownames(processed_data) <- paste(rownames(processed_data),
                                      processed_data$GeneNames,
                                      sep = "_")
    processed_data$GeneNames <- NULL
    anno_data <- anno_data()

    validate(
      need(!is.null(anno_data),
           message = "No groups assigned")
    )

    colnames(processed_data) <- anno_data$ID
    anno_data$axisLabels <- NULL
    if (input$calculateStats > 0) {
      f.df <- factor(anno_data$annotation)
      design <- model.matrix(~0+f.df)
      colnames(design) <- levels(f.df)
      fit <- lmFit(processed_data, design)
      f.df <- factor(anno_data$annotation)
      design <- model.matrix(~0+f.df)
      colnames(design) <- levels(f.df)
      fit <- lmFit(processed_data, design)

      cont.matrix <- makeContrasts(contrasts = input$hypoTestMat, levels =
 design)

      fit2 <- contrasts.fit(fit, cont.matrix)
      fit2 <- eBayes(fit2)

      return(fit2)

    } else {
      return(NULL)
    }
  })
```

To use the data from the statistical test automatically only one data frame is required from the output of limma, namely the coefficients data frame. By mapping the total comparisons to a value it is possible to allow users to traverse the different comparisons within the coefficients data frame by its index. Data frames can be constructed using this method to contain the fold change and q-value statistics for each comparison without knowing the comparisons explicitly. The number of data frames generated would be equal to the number of combinations calculated in the previous section and these data frames are stored in memory and this output is used to generate the final output graphs available in ProVision

| Uniprot ID | p-value | q-value | effect size | comparison | significant | Gene |
|---|---|---|---|---|---|---|
| Uniprot1 | 0.00001 | 0.001 | 3.5 | Wt vs Mut | Upreg | Gene1 |
| Uniprot2 | 0.00002 | 0.002 | -5.6 | Wt vs Mut | Downreg | Gene2 |
| Uniprot3 | 0.06 | 0.1 | -2.5 | Wt vs Mut | NS | Gene3 |

**Example of statisitical test output.**

```
statsOut <- reactive({
    fit2 <- statsTestedData()
    statComb <- statComb()

    #grab the data frame based on index
    d.out <- data.frame(ID = names(fit2$coefficients[,statsCycler$counter]),
                        pValue = fit2$p.value[,statsCycler$counter],
                        qValue = p.adjust(fit2$p.value[,statsCycler$counter],
                                          input$pvalAdjust),
                        EffectSize = fit2$coefficients[,statsCycler$counter],
                        comparison = statComb[statsCycler$counter])

    # Add significant column and calculate significance
    d.out <- mutate(d.out,
                    significant = ifelse(test = round(d.out$qValue,
                                                      3) < input$UserSigCutoff
                                         & d.out$EffectSize > input$UserFCCutoff,
                                         yes = "Upregulated",
                                         ifelse(test =  round(d.out$qValue,
                                                              3) < input$UserSigCutoff
                                                & d.out$EffectSize < (input$UserFCCutoff
                                                                      * -1),
                                                yes = "Downregulated",
                                                no = "Non significant")))

    # split the row names into uniprot and gene name
    d2 <- data.frame(d.out,
                     colsplit(string = d.out$ID,
                              pattern = "_",
                              names = c("UniprotID", "GeneName")))

    # remove redundant info
    rownames(d2) <- d2$UniprotID
    d2$ID <- NULL
    d2$UniprotID <- NULL

    return(d2)
  })
```

## Figure construction

### Volcano plots

Volcano plots are a variant of scatter plot and offer a rapid way to detect differences in data sets with replicates that have undergone statistical testing. It thus an effective method of visualising all the data found within a proteomics experiment both within comparisons and between comparisons. On such a plot, the effect size or fold change is plotted on the x-axis, a negative logarithm of the p-value is plotted on the y-axis and the points are coloured by q-value. Thus the greater the value on the y-axis the more significant and the greater absolute x value (can be either positive or negative) the greater the fold change. In ProVision the volcano plots are linked to the statistical test thus if there are two groups such as treatment control then you will only have one volcano plot. More than two conditions will create more volcano plots. The amount of volcano plots generated is the same as the amount of comparisons chosen during the statistical tests. Significant proteins can be labelled and are dependent on the statistical tests. Thus if there are no significant proteins it will not be possible to label something.

The volcano plot is constructed by determining the comparison first using the `statsComb` function and extracting the data from the statistical test performed. Addition of gene names is achieved by dynamically searching for the top N significant proteins provided by a slider on the front end and sorting the data frame based on this value. A new column is created for these genes and the altered data frame is used as the input data for the plotting system. The entire process is handled within memory. By keeping track of the index and the different comparisons users can change the plot that is displayed.

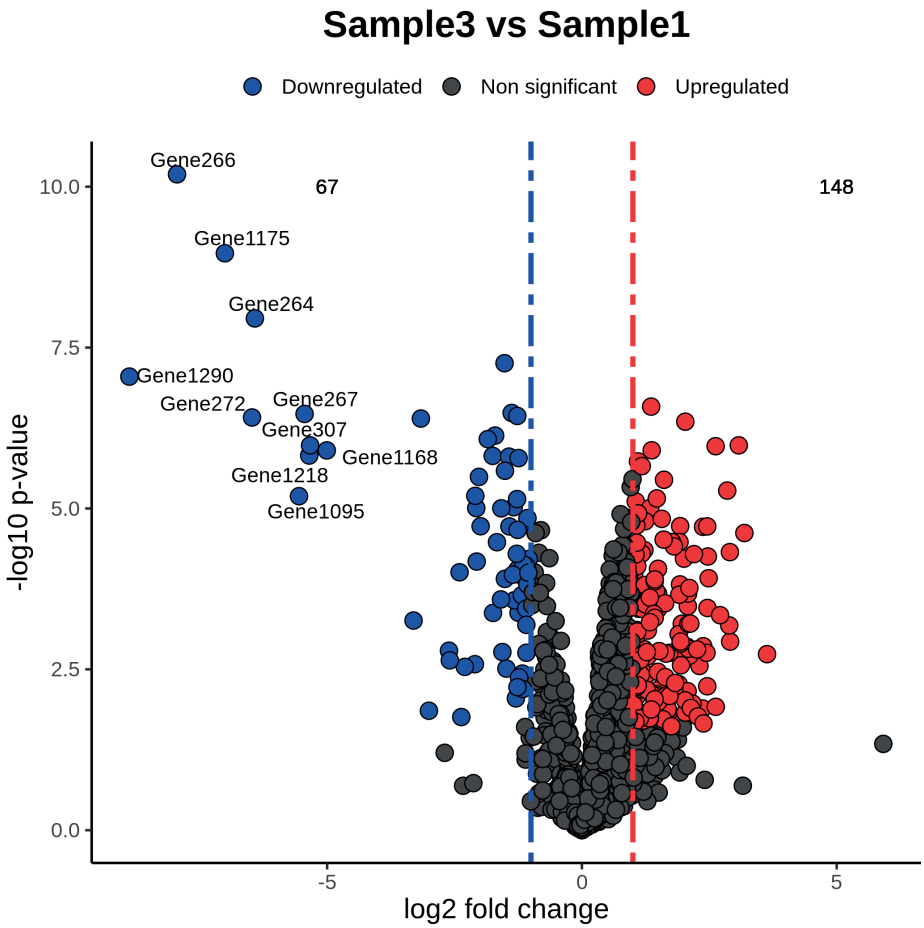



**Figure 7:** Example of a volcano plot representing the results of a statistical test. While p-values are plotted the points are coloured by q-value and cut-offs are determined by both fold change and q-value.

```
1.  voclplot <- reactive({
2.     d <- volcPlotData()[[volcCycler$counter]]
3.        p <- ggplot(d, aes(x=EffectSize, y=-log10(pValue), fill = sig)) +
4.          xlab("log2 fold change") + ylab("-log10 p-
value") + labs(fill = NULL) +
5.          ggtitle(label = input$volcTitle) +
6.          theme_classic(base_size = 14) +
7.          geom_point(pch = 21, colour = "black", alpha = input$volcAlphaChan
nel,
8.                     size = input$volcPlotPointSize) +
9.          scale_fill_manual(values=c("Non significant" = input$volcNS,
10.                                    "Downregulated" = input$volcDown,
11.                                    "Upregulated" = input$volcUp)) +
12.         theme(legend.position = input$volcLegendPostition,
13.               plot.title = element_text(face  = input$VolcTitleFace,
14.                                         hjust = input$volcTitlePos,
15.                                         size = input$VolcTitleSize),
16.               axis.title.x = element_text(size = input$VolcXSize),
17.               axis.title.y = element_text(size = input$VolcYSize))
18. return(p)
19. })
```

## Heatmaps

Heatmaps can be used to visualise the magnitude of differential expression across replicates and across groups, a metric which is lacking in the volcano plots. It is also possible to view the expression globally across the proteome using heatmaps. The data for the heatmap is either extracted from the statistical test or from the data frame prior to the statistical test depending on the heatmap generated. If a zoomed in heatmap is visualised then the data frame from the statistical test is sorted and annotated based on q-value and subsequently extracted to create a new data frame. The new data frame is passed to the heatmap plotting function below to be rendered on the front end. As with all the other data remains reactive with the rest of the data analysis work flow. While the statistical tests are used to zoom in to the significant proteins, these tests are not plotted on the heatmap and only fold change is represented.

```
UserHeatmap <- reactive({
    anno_data <- anno_data()
    if (input$generateHM > 0) {
      if (input$HMAllorSig == "All") {
        d1 <- allHeatMapData()
      } else {
        d1 <- HMPlotData()[[HMCycler$counter]]
        #d1 <- SigheatmapData()
      }
      if (input$HMdata == "averages") {
        colnames(d1) <- anno_data$annotation
        d2 <- sapply(split.default(d1, names(d1)), rowSums, na.rm = TRUE)
      } else {
        colnames(d1) <- anno_data$axisLabels
        d2 <- d1
      }

      validate(
        need(dim(d2)[1] != 0, message = "This comparison has no significant
 proteins"),
        errorClass = ".Shiny-output-error-validation {
        color: red;}")
      p <- pheatmap(d2, color = rev(brewer.pal(input$HMColChoice,
                                               n = input$HMcolScale)),
                    border_color = input$HMborderCol,
                    fontsize_col = input$HMColFontSize,
                    fontsize_row = input$HMRowFontSize,
                    angle_col = input$HMcolAngle,
                    cluster_cols = input$HMclustCols,
                    cluster_rows = input$HMclustRows,
                    clustering_method = input$HMClustMethod,
                    show_colnames = input$HMDispCol,
                    show_rownames = input$HMDispRow,
                    treeheight_col = input$HMColTreeHeight,
                    treeheight_row = input$HMColTreeHeight)
      return(p)}})
```

### Enrichment analysis

ProVision uses Webgestalt for enrichment analysis, specifically the associated R package (1). As the analysis is targeted towards proteomics data a lot of the inherent Webgestalt functionality has been stripped to streamline analysis. Two main enrichment tests are supported namely, over representation analysis and gene set enrichment analysis. From these analysis, either pathway enrichment can be executed or gene ontology enrichment. The available pathways are KEGG, Reactome, wiki pathways and Panther

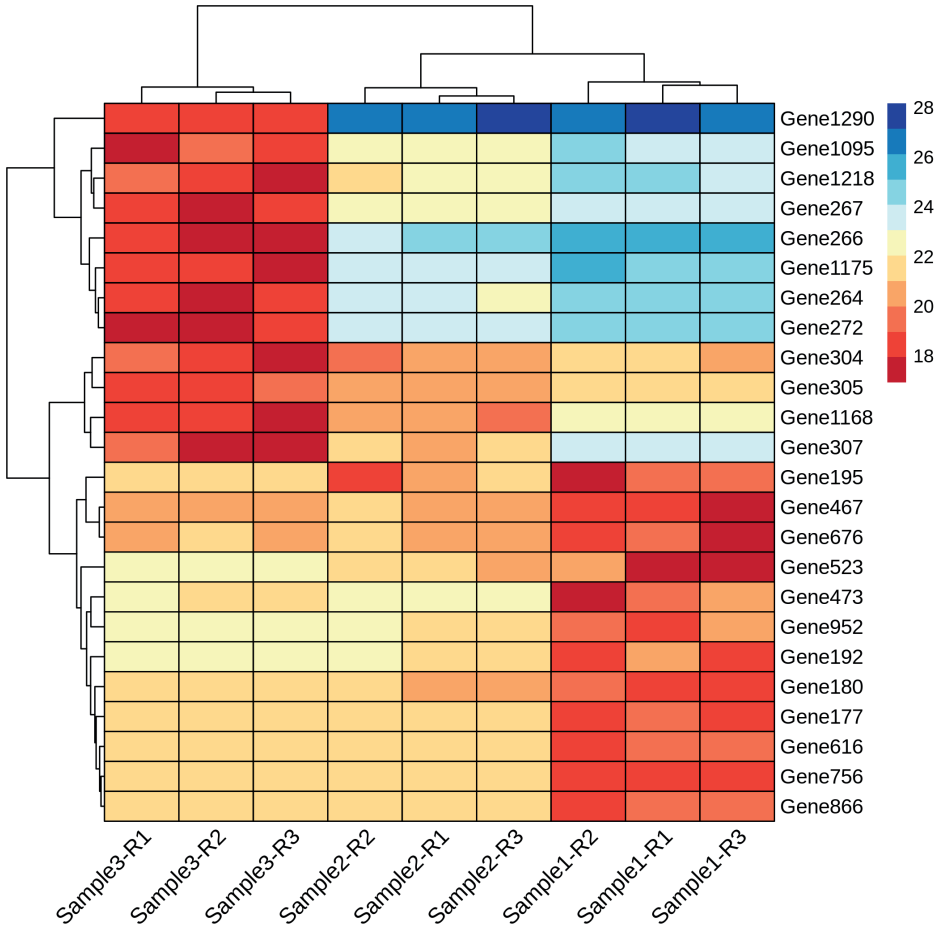



**Figure 8:** Heatmap depicting the top x significantly up- and down regulated proteins.

while the available gene ontologies are biological process, cellular component and molecular function.

Over representation analysis is a method that determines whether known biological processes are over-represented in a gene list (7). There are no values associated with over representation analysis and the enriched values represent either the up regulated or down regulated values as determined by an internal Fisher's exact test. Thus this type of analysis will find enriched gene ontologies or pathways within the given data set. An example use case would be to find the most enriched pathway in either a upregulated or downregulated set of protein identifiers. In the case of no significant proteins, the same analysis can be used on the protein identifiers to find information such as the most common cellular component found in a given data set.

Gene set enrichment analysis is related to ORA in that underlying biological processes or pathways can be enriched but takes into account a direct comparison by including the fold changes between two states. This is done automatically by ProVision by taking the previous statistical test into account. The gene set enrichment analysis first calculates an enrichment score that represents a number associated with the enrichment of genes at the top or bottom spectrum, i.e. up or down regulated. A statistical significance of the enrichment score is determined by permutation to produce a null distribution for the enrichment score. The P-value is subsequently determined by comparison to the null distribution and this p-value is adjusted for multiple hypothesis testing. Gene set enrichment analysis is useful in cases where fold change cut offs are arbitrary and an example would be to use all significantly different proteins and the fold changes in the enrichment. This can also be done globally without an hypothesis test to gain insight into proteomics data without the need for a formal test.



**Protein-protein interaction prediction.**

STRING is a curated database with known and predicted protein-protein interactions and detailed description of their software can be found in their latest manuscript (8). The STRING database has information on over 5 000 organisms regarding the interaction networks. In ProVision you can query the STRING database using either the up regulated or down regulated proteins that were determined by the statistical tests done previously. The data used to query the STRING database is reactive with the rest of the data in proVision, thus alterations that influence the statistical test will propagate to the STRING calls. It is possible that no information is displayed, this is due to no proteins being identified in the STRING network. In this case users can directly input Uniprot identifiers in ProVision to query the STRING database this way. It is important not to have any spaces between the Uniprot identifiers and each identifier must be

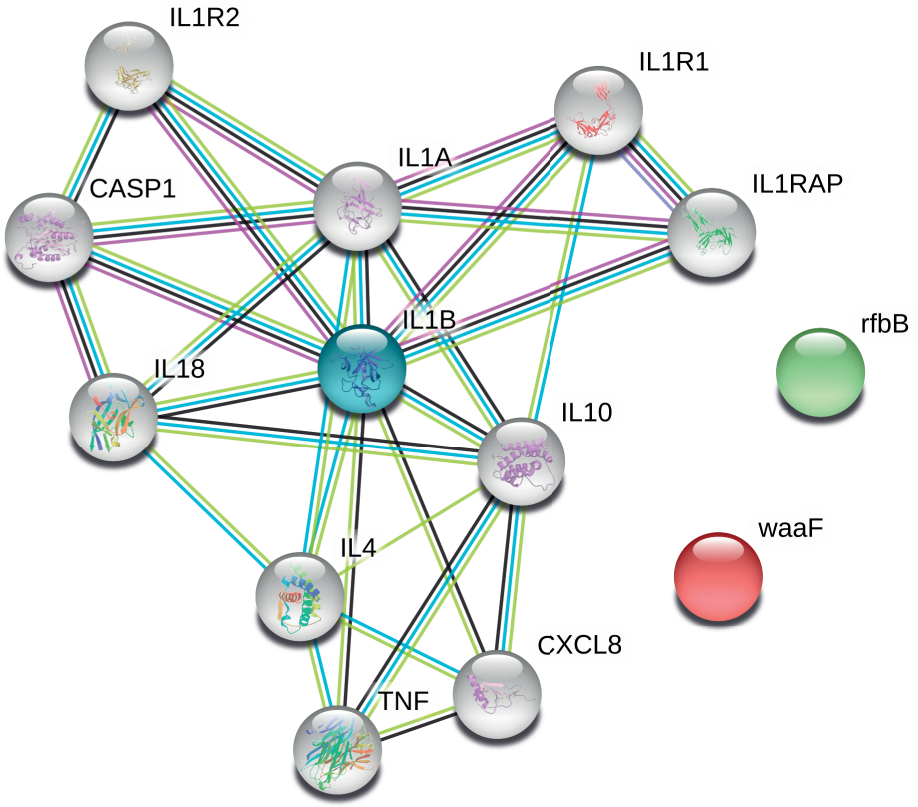



**Figure 9:** Example of the data extracted from the string protein-protein interaction application programming interface. Data is extracted automatically starting from the most upregulated – or downregulated proteins and submitted via an HTTP request.

separated by a comma for the query to work. Users can also opt to change the significant thresholds and the maximum nodes to fetch from the query. The significance threshold ranges from 0 to 1 000 where larger numbers represent stricter results.

A request is sent via http using a URL that is dynamically constructed from user input. Data is returned from the STRING API in JavaScript object notation and used to display data directly on ProVision without the user needing to navigate to the STRING website. To build the URL a list of protein identifiers are extracted from the data frame generated from the previous statistical test. These identifiers are filtered from most significant and decreases in fold change depending on the number of proteins added.

```
string_url_builder <- function(protein_query, sig_thresh, max_nodes) {
  validate(need(protein_querry > 0,
                message = "No proteins available"))
  if (protein_querry == 1) {
    data <- "network?identifier="
  } else {
    data <- "networkList?identifiers="
    protein_querry <- paste(protein_query,collapse=" ")}
  URL <- paste0("http://string-db.org/api/image/",
                data,
                protein_query,
                "&required_score=",
                sig_thresh,
                "&limit=",
                max_nodes,
                "&network_flavor=evidence")
  return(URL)
}
```

A few additional constraints are made available and the URL is built.

The base URL is built in the following way:

The *protein_query* parameter is built dynamically while *sig_thresh* and *max_nodes* are provided in the front end. To build the protein query the significant data is obtained first and dynamically handled based on user input:

Using the first function, the full URL can be constructed in an reactive expression to remain dynamic, there is also an option available to use custom inputs as well. In this case the data is not extracted from the statistics data generated earlier.

To render the image the URL is sent via HTML request and rendered using html tags in the application. As the URLis stored in a reactive STRING, the API call and rendered image will adapt to the user input.

## CONCLUSION

By creating this application we hope to assist those that are not adept at statistics and omics data analysis to reach rapid conclusions in their proteomics data without the need to study the field. Based on Shiny, the application is functionally written and quite large, thus not all code could be displayed within this document. The code is available on github for anyone to use, browse or distribute under an open source license.

## REFERENCES


1. Liao, Y. et al. (2019) 'WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs', Nucleic Acids Research, 47, pp. 199–205. doi: 10.1093/nar/gkz401.
2. Szklarczyk, D. et al. (2018) 'STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', Nucleic Acids Research, 47, pp. 607–613. doi: 10.1093/nar/gky1131.
3. Cox, J. and Mann, M. (2008) 'MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', Nature Biotechnology. Nature Publishing Group, 26(12), pp. 1367–1372. doi: 10.1038/nbt.1511.
4. Ritchie, M. E. et al. (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', Nucleic Acids Research, 43(7), pp. e47–e47. doi: 10.1093/nar/gkv007
5. Bantscheff M. et al. (2007). "Quantitative mass spectrometry in proteomics: a critical review". Anal Bioanal Chem 389(4):1017–1031
6. Diz P. et al. (2011) "Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work". MCP 10(3) doi: 10.1074/mcp.M110.004374
7. Boyle E. et al. (2004) "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". Bioinformatics 20(18):3710-3715 doi:10.1093/bioinformatics/bth456
8. Szklarczyk, D. et al. (2018) 'STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', Nucleic Acids Research, 47, pp. 607–613. doi: 10.1093/nar/gky1131.

# 8

# Summarising discussion

## APPLICATIONS AND LIMITATIONS OF MULTI-PLATFORM OMICS.

There has been a major boom in the technological advancement of molecular biology. This allowed researchers to pursue the molecular make up of living organisms using "big data" techniques such as genomics and proteomics. *M. tuberculosis* has benefited highly from this boom and has a large number of genomes sequenced which results in insights into the pathogen evolution and biology of this organism. Throughout this thesis genomics, proteomics and data driven approaches form the core around the studies conducted, along with *M. tuberculosis* as the primary target of investigation.

The genomes of a wide array of related organisms allows for the in depth investigation around topics such as molecular epidemiology, evolution, population bottlenecks and retention of genetic features across generations. However genomics and proteomics suffer from a reference bias, where all new sequences are compared to an established sequence. The established sequence is not necessarily the best sequence for comparison, especially in the context of *M. tuberculosis* where lineages and sub-lineages of the same species can differ with large polymorphisms. While using a reference sequence is sufficient for detecting cross species variation in relation to one another the variation is inherently only detectable if present within the reference. This creates a situation where the only features that can be detected are features that already exist. Depending on the specific use case this may not be a problem, such as determining the variation in single nucleotide polymorphisms across species, detecting novel features are lost in the process. Furthermore, the existing tooling is not sufficient to explore these lost characteristics and has thus been named the "dark matter" of the genome or the "dark genome". *De novo* assembly attempts to mitigate this loss by establishing a sequence in the absence of a reference through intra-contig alignment, using either *De Bruijn* graphs, overlap layout consensus or a combination of the two depending on the sequencing technology used (1,2). However, the use of short reads, such as those generated by the Illumina and Roche platforms, especially in highly repetitive genomes, is often unsuccessful or subpar. The *de novo* assemblies are often unfinished, especially in regions where *De Bruijn* graphs are prone to failure, such as the PE-PGRS and PPE-MPTR regions of *M. tuberculosis.* New technologies such as the PacBio long read sequencing can result in long contigs and often finish a small genome like *M. tuberculosis* but suffer from a higher error rate than the short read assemblers (2). As the genome of an organism forms the basis of genomics, transcriptomics and proteomics it is important to establish a sequence that is accurate to the organism of interest. An example where the reference of choice is of importance is seen with *M. tuberculosis*. Information regarding multiple large sequence regions will be lost in a transcriptomics

or proteomics experiment if the base reference was not from the same lineage as the subject. This is due to the large sequence polymorphisms that define each lineage which results in the inability to detect genes and in some cases full operons (3). The loss of this type of information is detrimental to the conclusions drawn, as these gain or loss mutations are themselves indicative of evolutionary adaptation to niche environments (4). In ideal circumstances a reference sequence should be a known parent strain of the test organisms, especially in proteomics where the sequence database is highly dependent on accurate information from the translation of reference genes.

Proteomics and transcriptomics is the natural continuation of genomics technology which aims to identify the functionality of proteins and/or transcripts within the organism. Both approaches are concerned with identifying properties associated with genes and their expression. Transcriptomics focuses on the mRNA transcripts of genes while proteomics is the study of the total proteins present. As we did not study transcripts in this thesis, the focus will be on proteomics and its applications. Proteomics is the comprehensive study of the protein make up of a cell and aims to generate maps of the total proteome signature. The complexity of the proteome in comparison to the genome is much more expansive than originally hypothesised and as more studies are conducted it becomes increasingly apparent that the cytosol and its regulation is a highly dynamic and complex environment with many interacting parts. It has become increasingly clear that we have indeed come a long way from the one gene one protein hypothesis (5). Proteomics has its strengths in providing one platform for a variety of different techniques, of which each technique can give a different insight into the biological mechanisms at hand. Using this technique with variation in experimental design in this thesis we could apply the following techniques throughout the various studies; differential quantification, post-translational modifications, protein turnover, temporal proteome response, spatial organisation of proteins. These techniques were applicable to both prokaryotic and eukaryotic organisms (6). Many more applications exist and new techniques are constantly under development, thus demonstrating the versatility and power of proteomics. Proteomics has its main limitation in the sheer complexity of the proteomics pipeline, where many individual steps are prone to error. Researchers that wish to perform a proteomics experiment generally require knowledge of the following: micro and/or cell biology to grow cells to usable quantities, molecular biology to extract proteins and perform quality checks, analytical chemistry to prepare and separate the samples, working knowledge of the concepts involved in tandem mass spectrometry and mass analysers for proper experimental design as well as working knowledge of statistics and programming capabilities to analyse the large amounts of data. For the best results in proteomics the one responsible for the project design and original hypothesis needs to keep all these steps in mind when designing

the experiment. Proteomics also suffers greatly from a reference bias, while genomics and transcriptomics have made major strides in *de novo* assembly (7–9), proteomics is still lacking in this regard (10). Database search algorithms used in matching real tandem mass spectra with theoretical spectra assume that all spectra is present within the database and that the masses are correct for each spectra. This is however not the case when the theoretical database is obtained from the sequencing of a reference sequence that is dissimilar to the study organism. This also poses the additional challenge of post-translational modifications, where these are lost if not explicitly added to the search parameter. These modifications can also occur after cell lysis, a common modification is oxidation. Thus many proteins that are present within the total protein content are not found due to unknown variable modifications on the proteins. In addition, peptides with trypsin missed cleavages will also be disregarded in the search unless specifically stated otherwise, which in turn significantly increases the database search time. With the knowledge of all possible post-translational modifications and including all possible missed cleavages as well as extreme computing power can mitigate these effects. This is however of no consequence if the protein and its theoretical spectra are not in the sequence database to begin with. The way to ensure that all real spectra have a theoretical counterpart is by ensuring that the correct reference sequence is used. To achieve this, researchers have merged the fields of genomics and proteomics to create proteogenomics (11).

Proteogenomics aims to address the reference bias and allows for a broad range of detection. Peptides may also be lost to database search algorithms due to mutations in the genome of even slightly unrelated organisms or due to complete lack of operons as mentioned earlier. Approaches to identify peptides with that may not be present is using sequence tag database searching or *de novo* peptide assembly, both of which are error prone and fail for large scale studies (12,13). Proteogenomics bypasses these restrictions by applying the established algorithms used to finish genomes to a specific organisms of interest. The proteins of the specific organisms are determined from the reading frames and this is used to establish the proteome database. Using this method, especially if the assembly of the target organisms is a complete or near complete genomic *de novo* assembly, all the mutations and coding regions are present and can be used for inference. This technique has allowed researchers to identify a large number of novel peptide variants and is further extended to identify chimeric proteins in *M. tuberculosis* as discussed in chapter 4 (14,15). Proteogenomics does have one obvious draw back when considering large amounts of data. As inference of a peptide is based on a statistical test and that statistical test is controlled by a false discovery rate, the larger the database becomes the stricter the test will become and thus real peptides and by extension proteins will be lost.

As technologies become cheaper and advances are made in the software available to researchers, studies may incorporate genomics, proteomics and indeed proteogenomics for targeted studies in organisms like *M. tuberculosis*. In chapter 4 we applied proteogenomics by using Illumina short read sequence data and proteomics to identify novel gene fusions as well as functional chimeras in *M. tuberculosis* clinical isolates. As mentioned a full genome is required but a reference introduces the reference bias and thus a *de novo* assembly is needed, however the *de novo* assembly produces an unfinished genome. In order to bypass this we used reference assembly to map the majority of reads and created custom software to extract regions around breakpoints that span multiple genes. These regions were subsequently used for *de novo* assembly for a more targeted approach and single base resolution of the breakpoints (see Chapter 1, figure 2). The gene fusion resulting from the combination of two distally located genes was thus determined and translated to identify potential chimeras. While this approach was suitable for the identification of genomic elements that fit this criteria, the use of PacBio sequence data and long contigs or full *de novo* assemblies can greatly increase the discovery of novel genetic elements in important pathogens such as the *M. tuberculosis.* As the cost of long read sequencing decreases, the use of this type of sequencing in combination with proteomics would be highly beneficial and may become the mainstay and surpass Illumina as the default platform, especially with regard to *M. tuberculosis*.

## GENETIC LOSS IN THE MYCOBACTERIA AND THE ROLE IT PLAYS IN MYCOBACTERIAL SURVIVAL.

As an intracellular pathogen, *M. tuberculosis* is prone to genomic decay (16). While there is relatively low genomic diversity in comparison to other organisms there is a high amount of recombination events, large scale polymorphisms and single nucleotide polymorphisms to drive genomic change (17). One of the more striking features of *M. tuberculosis* is the lack of horizontal gene transfer to drive evolution (18,19). This is a unique position for an asexually reproductive organism as new DNA and by extension new functions need to be created from existing genetic material. This phenomenon explains the high prevalence of single nucleotide polymorphisms as the resistance causing mechanism used by *M. tuberculosis* to combat antibiotics, while other bacteria often transfer resistance genes via plasmids (20,21). The large sequence polymorphisms are of interest with regards to mycobacterial evolution. These large deletions are used to stratify species into various sub-lineages which are highly geographically restricted (Figure 1). Typically genetic decay occurs when a symbiont provides a function to the host and in turn the host assumes functions for the symbiont. As *M. tuberculosis* is

parasitic, the host is actively attempting to eliminate foreign the organism and thus the pathogen is under evolutionary pressure to survive. Therefore the fact that there is a geographically restricted signature is a strong indicator of direct competitive evolution through the loss of genetic material with each sub-group of hosts (Figure 1) (22) .

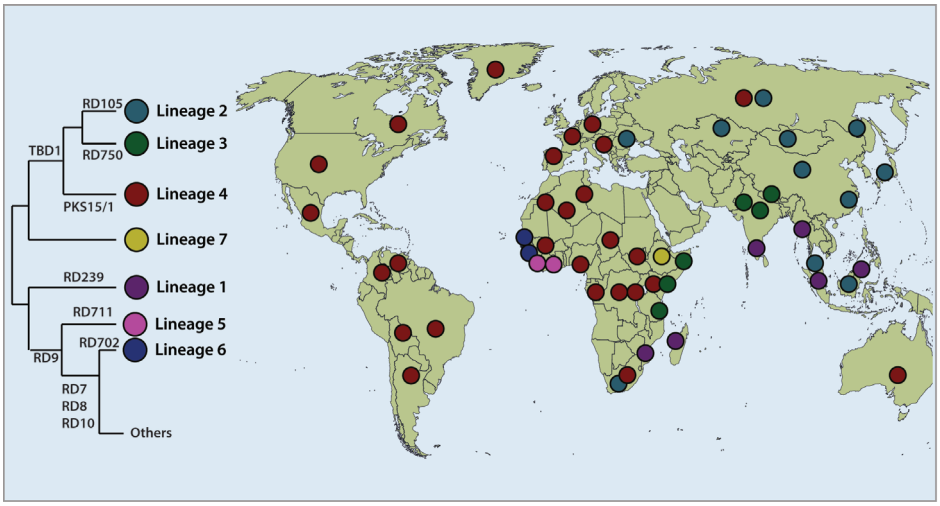



**Figure 1:** large sequence polymorphisms and global distribution of M. tuberculosis isolates adapted from Coscolla & Gagneux 2014. The figure demonstrates the differentiation of M. tuberculosis isolates in isolated pockets across the globe, indicating competitive co-evolution with the host.



In chapter 4 we have demonstrated the functionality of genetic loss through the creation of gene fusions and that these genetic features are able to form functional proteins. This was achieved by use of proteogenomics and accurate break point prediction by combining reference and *de novo* assembly in regions with multi-gene deletions. This same phenomenon has been demonstrated for large sequence polymorphisms used for species stratification as well (chapter 4). The genetic loss is thus likely a major adaptive trait of *M. tuberculosis* to its specific niche and used for the *de novo* biosynthesis of genes from two related or unrelated parent genes, thereby greatly increasing the utility of a shrinking genome. As mentioned earlier, these features require at least lineage specific reference genomes to uncover but also provides a challenging aspect in the use of proxy organisms to study *M. tuberculosis* phenotypes observed across lineages or species. As an example, *M. bovis* BCG represents a sub-optimal model for *M. tuberculosis* study as it lacks the RD-1 region, which is responsible for a large subset of mycobacterial virulence (23). Another may be found in the lineage two strains of *M. tuberculosis,* these strains are more virulent and prone to higher transmission, which is interesting as their defining factor between lineage four strains is the loss of genetic material (24,25) . The deletions responsible for increased functionality may not necessarily be known

stratifying deletions but also more spontaneous variation. The loss of the *ppe38-ppe71* region blocks secretion of the PE-PGRS proteins and causes an increase in virulence over time (26). This region is intact in most, but not all lineage four strains, and can be reformed even after deletion as to once again secrete the PE-PGRS proteins (see Chapter 4). As mycobacterial research continues into the omics era, the large sequence polymorphisms and other less obvious changes in the genome may provide new and interesting information in the understanding of *M. tuberculosis* evolution.

## CONSTRUCTING A HIGHLY VERSATILE MODEL FOR M. TUBERCULOSIS.

To control the variation in the different *M. tuberculosis* strains, *M. tuberculosis* H37Rv has become the standard reference strain. Often models for pathogenic mycobacteria are used for investigating the pathogen for safety and ease of use reasons. However, the main problem with using models for investigating a specific organism is that the genetic composition is naturally different to that of the target. This can be an issue for *M. tuberculosis* as the use of a non-pathogenic mycobacteria often defeats the purpose. In addition, a less pathogenic mycobacteria is while not deadly is still unsafe and genetically distinct from their pathogenic counterpart. Models are however necessary for the study of tuberculosis as a BSL-3 is a very restrictive environment that often is not capable of supporting advanced machinery or execution of advanced experimental techniques. Auxotrophic variants of *M. tuberculosis* provides a solution to this problem directly by creating minimal genomic variation while creating safe and widely applicable proxy organisms. By creating a safe alternative to *M. tuberculosis* and decreasing the biosaftely level from three to two greatly increases the calibre of experimentation possible with such an pathogen (see chapter 2).

In chapter 2, we describe the characterisation of an auxotrophic *M. tuberculosis* strain, lacking *leuD* and *panCD* which renders this *M. tuberculosis* H37Rv strain incapable of growth in absence of leucine and pantothenate. We show that the auxotrophic *M. tuberculosis* provides the means to study diverse applications and has similar physiological responses as *M. tuberculosis* H37Rv. The *ΔleuD/panCD* strain also has a defined genetic background with relation to the parent and by avoiding excessive passaging the genome remains stable. It is also possible to use such an organism to mimic the deletions or single nucleotide polymorphisms present in other lineages of *M. tuberculosis*. This is particularly useful for testing drug resistance-conferring mutations directly in a *M. tuberculosis* background without creating a more dangerous version of the pathogen first.

The leucine deficiency of the auxotrophic strain can be exploited and used for quantitative whole proteome labelling experiments by use of stable isotopes of amino acids in cell culture (SILAC). SILAC is well established and typically involves total proteome labelling of Leucine and Arginine, as these are both essential amino acids and the targets of the trypsin protease. Thus the use of SILAC and proteomics is prevalent and used in this study to label human proteomes as well. By using heavy leucine we were able to label the proteome of *M. tuberculosis.* This extends the usability of the *M. tuberculosis leuD/panCD* mutant immensely and enables the use of advanced quantitative mass spectrometry techniques such as detection and quantification of post-translational modifications, highly accurate measurements of protein abundance, protein turnover measurements and more (see chapter 1). This greatly extends the reach of this specific auxotroph and is arguably a contender for the *de facto* model of *M. tuberculosis* when the pathogenic counterparts are out of the question.

## AN EXTENSIVE QUANTITATIVE PROFILE OF GLOBAL M. TUBERCULOSIS PROTEOME REGULATION DURING ACID STRESS.

We used a SILAC labelled *M. tuberculosis ΔleuD/ΔpanCD* mutant for various quantitative proteomics experiments, including differential protein abundance, quantitative post translational modifications and protein turnover. A striking observation was made in the protein turnover experiments with the auxotrophic *M. tuberculosis* strain. This technique measures the differential protein content over time and is tracked by stable isotope labels. This is achieved by either growing the target organisms in a heavy isotope and fully incorporating this isotope and pulsing the organisms with the light or natural version. This technique can also function using the heavy isotope to pulse the organisms and this pulsing typically follows and physiological stressing event that will demand protein synthesis. In this example, the new proteins that are synthesised contains the light version of the amino acid as this theoretically the only amino acid present, there may however be some heavy amino acids present that are recycled from older proteins (see chapter 5, supplementary figure 3A). As the organisms equilibrates to the new environment the proteins will gradually be turned over until the organisms doubles and the proteins are diluted by half. All of these processes can be calculated using a first rate reaction equation that takes into account dilution.

In chapter 3 we used the auxotrophic *M. tuberculosis* strain characterised in chapter 2 to investigate the mycobacterial response to acid stress. We specifically investigate the global proteome profile, including differential abundance, quantitative phosphoryla-

tion and protein turnover using SILAC and proteomics. At the physiological pH, a number of proteins with the light version of leucine were present and the protein turnover was calculable. However, little to no light leucine was detected in the proteomes of the bacilli pulsed in acidic pH, thus rendering the calculation of half-lives void under these conditions. In retrospect, it is likely that the acidic pH depolarises the membrane enough to block the amino acid uptake from the environment. Thus during this condition, extracellular leucine is not incorporated into the proteome and not detectable by mass spectrometry. It is tempting to speculate whether the same lack of amino acid transport occurs when subjected to the acidified macrophage environment. However, all of the components associated with negating macrophage acidification are in place and thus the effect may be less pronounced. This is nonetheless an important observation to consider when using a model organism, while all genetic and phenotypic traits are controlled for there are still room for surprises with every experiment.

In the acid-free control sample the protein turnover of *M. tuberculosis ΔleuD/ΔpanCD* did provide information on the general state of proteome homeostasis during passive growth. By dividing the protein half-lives into categories, namely class I (fast) class II (intermediate) and Class III (long lived) we were able to elucidate the broader functions of the proteins represented in each class using enrichments (see chapter 3). The intermediate class had the most interesting observations, with bacterial responses to the host and the type VII secretion featuring prominently. These included the ESX conserved components of ESX-1, ESX-3 and ESX-5. We also observed increased phosphorylation of type VII secretion factors during acidic stress. As has been shown with the sequencing of *M. bovis* BCG as well as routine passaging, the virulence of a pathogen is lost over time. It is however clear that the type VII secretion systems are active in optimal growth conditions without the need for stressors. Our results further indicate that culture filtrate protein 10 is phosphorylated at a greater propensity when upon exposure to acidic stress. The phosphorylation of a protein is associated with a form of signalling and it is tempting to speculate whether there is some form of signalling occurring with this virulence factor or whether phosphorylation mediates conformational changes to create an active CFP-10. Indeed, at least partial blockage of acidification is a pre-requisite for phagosomal escape and phosphorylation patterns on the effector molecules may be the method of interacting with the environment (27). It is however more likely that the increased phosphorylation is due to more EccA and EccC ATPase proteins releasing a free phosphate that binds to proteins like CFP-10 when exposed to stress (28). This would have been confirmed if the protein turnover of the ESX secretion systems stayed constant during acid stress but unfortunately we were not able to measure this.

## THE *PPE38-PPE71* DELETION IN CIRCULATING *M. TUBERCULOSIS* STRAINS AND ITS CONSEQUENCES.

Since its discovery in the mycobacteria, the type VII secretion system has become a mainstay of mycobacterial research especially in the context of virulence. These systems are conserved in the mycobacteria and found in other gram positive bacteria as well. In the pathogenic mycobacteria there are a total of five type VII secretion systems designated ESX-1 to ESX-5, where ESX-2 and ESX-5 are the most recent and specifically of interest is ESX-5 which only present in slow growing mycobacteria, including most pathogens (see chapter 1 for more details). In the context of virulence and type VII secretion systems, arguably the ESX-1 and ESX-5 secretion systems are the most important for virulence. The ESX-1 secretion system is well documented in its role in virulence and is most notably involved in blocking phagolysosome maturation (29–31) and has been reviewed extensively (32–34).

The ESX-5 secretion system is unique to the pathogenic mycobacteria and responsible for the secretion of the PE-PGRS and PPE-MPTR proteins to the extracellular milieu. The ESX-5 system and the roles of the PE/PPE proteins have also been reviewed recently (35). The PE-PGRS and PPE-MPTR proteins provide an intriguing avenue of research in their own right. As *M. tuberculosis* is subject to reductive evolution as a result of an intracellular lifestyle, retaining certain coding regions above others should signify an evolutionary advantage due to the natural selection at play (16). The PE-PGRS and PPE-MPTR proteins are clonal expansions which contribute approximately 10 percent of the total coding potential in the *M. tuberculosis* genome. They are also unique to the pathogenic mycobacteria and secreted through the ESX-5 system, which is in its self also a unique type VII secretion system (34). Furthermore, the surface localisation of these proteins strongly suggest some form of interaction with the host cytosol and cellular processes (36). The PE-PGRS proteins are also highly immunogenic with individual members implicated in diverse roles most of which are related to survival in harsh environments such as in macrophages and/or granuloma's (37–39). Taking the evolutionary context, the genetic features and the high immunogenicity factor into account, one of the first theories on the collective function of these proteins suggested a role in antigenic variation and thereby escape host responses to infection (40,41). This has however fell out of favour due a lack of recombination of these proteins within the host. There is also a general lack of evidence for this phenomenon and it is more likely that each of these proteins mediate a distinct function (39). The discovery of the *ppe38-ppe71* operon and its role in modulating secretion of the PE-PGRS proteins to the extracellular milieu provided striking insights into the working of these proteins (26). Namely, that the PE-PGRS proteins are controlled in bulk by a single operon, the

deletions of this operon is found in clinical isolates of tuberculosis and this deletion results in increased bacillary load in mice (26).

In chapter 4 we used genomics and proteomics to investigate the prevalence of this deletion as one of our prominent examples of interesting multigene deletions in *M. tuberculosis*. We found a greater frequency of the *ppe38-ppe71* deletion in the lineage two isolates of *M. tuberculosis* above lineage four strains. Interestingly, the lineage two strains are associated with an increased pathogenicity profile and higher mutation rates in comparison to the lineage four counterparts (42,43). It is tempting to add the *ppe38-ppe71* deletion to the list of contributing factors. This is however ill advised without more detailed investigations, given the complexity of any given organism and how little is actually consolidated about the molecular dynamics of even simple life forms. By changing perspective slightly, the correlation between the *ppe38-ppe71* deletion and the lineage two strains is telling as the lack of PE-PGRS proteins does not seem to hinder these strains in the infectiveness and transmissibility. Our investigation into the lineage four *ppe38-ppe71* showed that the breakpoints do not always form in the same manner, and in some cases the breakpoints can recreate *ppe38* as a single gene from the original operon and thus forming a gene fusion. We further demonstrated that this gene fusion forms a functional protein able to secrete the PE-PGRS proteins once more (chapter 4). It would be interesting to investigate if the lineage four strains that have the reformed *ppe38* compared to the full deletion have an increased virulence and mutation rate or if the virulence remains stable, and *vice versa* for the lineage two strains that do not have the *ppe38-ppe71* deletion. This may give indication into the role of this deletion within the circulating isolates and whether the lack PE-PGRS proteins rather promote a more aggressive phenotype. Highly immunogenic proteins also function to alert the immune system to the presence of the bacilli. The lack of these triggers can provide valuable time for the bacilli to establish an infection before being subject to the full force of the immune system. Additional time to response, especially for a slow growing organisms could mean the difference between eradication and persistence/active infection. Indeed, the increased hypervirulence seen in mice was only evident at later time points through the presence of bacterial load (26). As bacterial growth is exponential, it is likely that additional doublings at an early stage can result in a greater overall population over a few generations compared to bacilli that were curbed by innate immune responses.

## INFLAMMATORY RESPONSES TO *M. TUBERCULOSIS* WITH AND WITHOUT PPE38.

The presence of the *ppe38*-ppe71 deletion presents an anomaly with regard to *M. tuberculosis* physiology in the context of competitive co-evolution. Natural selection is a slow march towards an optimised state. Thus, as favourable mutations rise, the bacteria gain fitness and the mutation propagates through the population. Parasitic reductive evolution suggests that only the most favourable traits, which provide a competitive edge, would propagate in an organism that loses their genetic material over time. Thus the coding regions that are present should be beneficial as genetic "real estate" is at a prime rate within these organisms. The *ppe38-ppe71* deletion challenges this dogma to an extent. While the genes themselves are present, their protein products and the effects they may have is functionally abolished. The most likely point of interaction between the PE-PGRS proteins and the host would be during early infection or active infection, as these proteins are secreted and the Type VII secretion system is active during typical growth states (see chapter 5). During dormancy or persistence, the bacilli has a low metabolic state and thus using active transport is likely at a minimum (44). Furthermore, we speculated that an early advantage may result in increased bacterial growth over time which lends an explanation to the increased virulence observed (26) .

In chapter 5 we investigated the human macrophage response to infection by *M. tuberculosis* with and without the *ppe38-ppe71* deletion. The PE-PGRS and PPE-MPTR proteins represent a broad category of proteins which have the potential to elicit a wide array of responses. To address this we used a discovery based proteomics approach to measure the differential abundance of proteins over time. We also used SILAC labelling of the THP-1 macrophages over time to calculate the protein turnover in response to infection. The major findings observed in chapter 5 included a low response to infection with a marked decrease in inflammation when the macrophages were challenged by *M. tuberculosis ppe38-ppe71* mutants. This was validated by Western blotting and we further investigated the immune signalling pathways involved with immunofluorescence against the Nuclear factor kappa B (NF-kB) sub-units. We found that the signalling of this pathway is altered in macrophages infected with *M. tuberculosis* isolates deficient in PE-PGRS/PPE-MPTR secretion. Similar phenotypes have been observed using *M. marinum ppe38* transposon mutants, while cytokine secretion using *M. tuberculosis* using dendritic cells as the host had only minimal differential regulation (45,46) (see chapter 5 for details).

At first glance a decrease in macrophage response seems intuitive, as fewer virulence factors are available to cause a stimulation of the immune receptors. Therefore less

signal is propagated and less immune responses are observed (figure 2). However, the inflammatory response decreased over time indicating that the macrophages are actively decreasing the response to infection, a phenotype that was reflected in both temporal proteome abundance and protein homeostasis. Enrichment of the major canonical pathways suggested a regulation in NF-kB signalling. This pathway is central to controlling fundamental immunological responses and controls critical processes such as inflammation. Alterations in NF-kB can lead to fatal disorders either through runaway inflammation or abolishing the innate immune response and thus resulting in a susceptibility to infection (47). At its core, the NF-kB pathway is a transcription factor pathway that is initiated by receiving signals from the environment and acts upon these signals by transcribing hormone like proteins, i.e. cytokines (48). These cytokines are the basal effectors of the immune system which can either stimulate the full activation of the immune system, decrease the immune response, cause cells to lyse or differentiate into other cell types (49). By modulating the NF-kB pathway, a pathogen can alter the host on a fundamental level and thus secure its survival indefinitely. The mechanism of NF-kB translocation directs the type of response that can occur. This is typically split into the canonical and non-canonical pathways. Both of these can control inflammation by either rapid transcription of inflammatory cytokines or a delayed response if the non-canonical pathway is stimulated (50). We determined

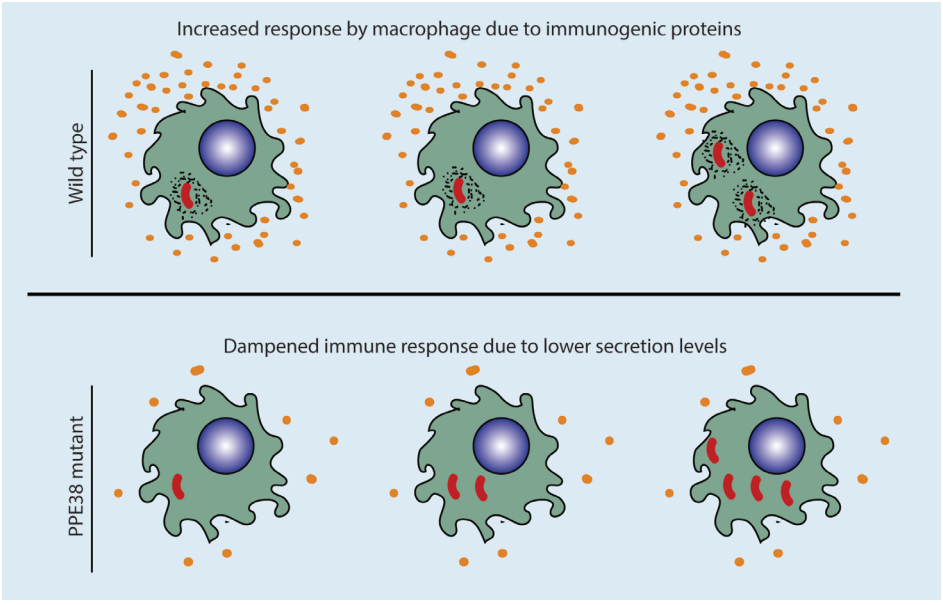



**Figure 2:** Macrophages infected with wild type or ppe38-ppe71 knock out tuberculosis. The ppe38-ppe71 mutant stimulates a lower overall immune response compared to wild type. Over time, this response may result in a greater intracellular bacterial load compared to wild type.

the extracellular protein profile of *M. tuberculosis ppe38-ppe71* mutants in comparison to wild type prior to infection in *in vitro* growth media. This gave indication on the PE-PGRS and PPE-MPTR proteins involved in PPE38 mediated transport across the cell. We found a clear cluster of highly regulated PE-PGRS and PPE-MPTR proteins which are likely the drivers of the phenotypes we observed *ex vivo* (see discussion of chapter 5 for more details). The components of NF-kB can combine in multiple ways and each permutation can have a major effect on the cytokines that are transcribed. This translocation was directly investigated and alternative NF-KB pathways associated with anti-inflammatory responses were found. Specifically, when infected with the *ppe38* mutant, RelB translocation occurred instead of RelA, strongly indicating an anti-inflammatory phenotype (Figure 3). This discrepancy indicates a molecular switch controlled by *M. tuberculosis* through influencing translocation of Rel sub-units. This molecular switching of inflammatory pathways has been observed before, where *Salmonella* SPI-II systems mediate switching between M1 and M2 macrophages thereby reprogramming the macrophage signalling pathways (51). A similar mechanism may

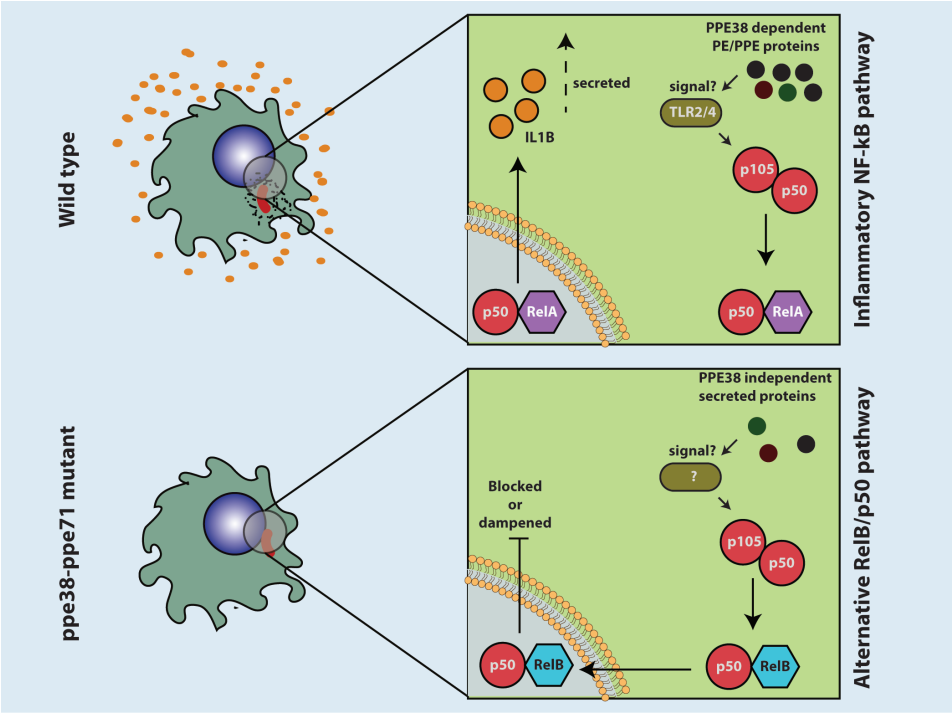



**Figure 3:** Schematic illustration of the differential Nuclear factor kappa B signalling pathways identified in wild type and ppe38-ppe71 deficient mycobacteria. The lack of PPE38 dependent PE/PPE proteins stimulates a p50/RelB alternative pathway while the wild type signals through canonical NF-kB pathway. The canonical pathway stimulates secretion of IL1-B and creates an inflammatory response, the p50/RelB has been shown to dampen the inflammatory response.

be at play with the use of *ppe38* to modulate PE-PGRS and PPE-MPTR secretion which in turn can selectively stimulate the immune response thereby achieving a type of re-programming. The lower immune response also provides the evolutionary imperative to remove the secretion of these proteins as discussed above. The lower the immune response to infection the more time to establish an infection. It is curious that the mechanism of secretion is lost but not the individual PE-PGRS proteins. Perhaps *M. tuberculosis* isolates with the *ppe38-ppe71* deletion still synthesize these proteins and release them upon lysis to interact with other bacteria that are still intact. Thereby maintaining the overall functionality of these proteins, even in the absence of PPE38 to mediate the secretion. Recent findings of a protease, PecA, was shown to cleave the PE-PGRS proteins with a possible cleavage site located in the PE region (52). This indicates that the variable PGRS part is free from the cell and able to function independently. While a rather extreme hypothesis, there is some imperative for a lysed mycobacterial cell to release the PE-PGRS proteins to the extracellular milieu and still interact with the host.

To summarise, there is evolutionary imperative for the deletion of the *ppe38-ppe71* operon in circulating clinical isolates (chapter 4) (26). These deletions likely contribute to a lower early immune response due to the lack of immunogenic proteins on the surface, thereby providing a more favourable environment. The lack of immune response likely translates to increased bacterial load as the bacilli double over time as seen previously (26). Thus bacilli lacking the *ppe38-ppe71* deletion and by association the secreted PE-PGRS and PPE-MPTR proteins allow these bacilli to trade an early virulence phenotype for a late virulence phenotype. The infection profile of *M. tuberculosis* is however complex and cannot only be categorised by the macrophage infection profile. Alveolar macrophages are for one more permissive than others and thus may have a more pronounced phenotype when infected with the *ppe38-ppe71* deficient macrophages. Furthermore, *M. tuberculosis* is characterised by the formation of granulomas at the site of infection (53) and PE-PGRS proteins have been observed in *M. tuberculosis* granulomas (54,55). Many studies approach the PE-PGRS and PPE-MPTR proteins from the angle of the bacilli with the question focusing on the function of these proteins. It may however be more fruitful to shift focus towards the modulation of the immune system by these proteins in order to decipher more of this unique physiology.

## INVESTIGATING *M. TUBERCULOSIS* RELAPSE THROUGH GLOBAL LUNG PROFILES USING MASS SPECTROMETRY.

In the Western and Eastern Cape provinces of South Africa, the lineage two strain is the most common strain infecting tuberculosis patients (56). As mentioned, this lineage presents the highest incidence of drug resistance and is the most aggressive (read pathogenic) (42,43). Unfortunately the completion of treatment is only at 53 % and among these there is a higher mortality rate associated with the drug resistant and aggressive tuberculosis isolates (57). Completion of treatment is of utmost importance, as *M. tuberculosis* can persist during the treatment process and later re-emerge to infect the individual once more. However, the mean relapse rate remains at ~8 % (58–60). The lung is the largest surface in contact with the outside environment and thus highly trafficked by immune cells to clear any particulates or organic agents inhaled by the individual (61). The lung profile and how the proteins present during infection with these isolates is a untapped resource.

In chapter 6 we procured sputum and bronchiolar alveolar lavage (BAL) fluid from patients at the start of tuberculosis treatment and at the end of treatment in the Western Cape regions of South Africa. These patients were predominantly infected with the *M. tuberculosis* lineage two strain (data not shown). The patients were also subjected to fluorodeoxyglucose positron emission tomography/computed tomography (FDG PET/CT) scans to classify lung lesion activity. We found that approximately 10 % of patients retained live *M. tuberculosis* at the end of treatment and two patients relapsed. Proteome analysis of the BAL fluid could differentiate between active tuberculosis cases and end of treatment cases. We also found that the relapse case demonstrated a BAL fluid proteome signal similar to the active tuberculosis cases at the end of treatment. The sample set is however too small to draw any real conclusions and this specific relapse sample without further investigation and may not represent a total population. The rarity of relapse cases and the invasiveness of extracting BAL makes investigating the in lung relapse profile extremely challenging. If this can be overcome, a greater proteomics BAL study with sufficient numbers in active, end of treatment and relapse cases has great potential. Identification of target profiles or circulating molecules within these patients can indicate the effectiveness of treatment, whether treatment can be stopped early or should be extended.



### Creating an open source proteomics data analysis platform.

Throughout the studies presented within this thesis, proteomics forms a central point. The analysis of proteomics data, and indeed other large dataset manipulations is still an open field with active development and there is no one-fits-all method. Certain para-

digms may be followed but this can result in sub-optimal results that stem from the data handling rather than the actual biological experiment. In many ways the analysis of the data should be decided in the context of the research question but is often disregarded up until the raw data is generated. The major steps in the data analysis of a proteomics experiment includes filtering of unwanted data, transformation, filtering of biologically non-sensical data, imputation, statistical testing and visualisation. Each of these steps require user input to some extent and may be of use in certain cases but irrelevant in others. It is up to the researcher to determine these factors, but the sheer volume of data often causes traditional methods such as spread sheets to be cumbersome. There are software packages available, with the *de facto* analysis platforms being Perseus, which is free, and Proteome Discoverer (62). Perseus attempts to create methods for the general analysis of proteomics data but without prior knowledge in the workflows a potential newcomer is bombarded with many hours of tutorials resulting in a high skill ceiling. Furthermore, the graphics quality is lower than other platforms and limited in its customisability within Perseus. The R programming language offers the most flexibility but the skill ceiling is even higher as it requires users to be knowledgeable in software engineering paradigms as well as implementation of the various steps. The power of R is however unparalleled and any analysis can be easily tailored to a research question retrospectively and quality graphics can be generated, speeding up the analysis. The Bioconductor and CRAN libraries of R also provide a strong community with many developments in proteomics to facilitate the analysis (63,64).

In chapter 7, we created a cloud based proteomics data analysis platform in the R language for the analysis of common proteomics experiments which we called ProVision (Proteome Vision). The main driver behind creating this software was to rapidly analyse common proteomics data with the freedom to tailor the analysis to the research problem while retaining the full functionality provided by R. As we performed many different types of analysis it was imperative that we created a flexible working environment. The alternative was to create a R library that performs similar operations as discussed here, this was however not as user friendly as using a R back end with a graphical front end for quick analysis. We achieved the low skill ceiling and rapid prototyping by creating a clear and intuitive workflow with control over each section. An important feature implemented is how the data flows through the platform, where changes created in the first step are propagated downstream to the figures. This allows for rapid experimentation of data analytics techniques to achieve the most optimal biologically relevant results. Proteomics data analysed with ProVision makes use of MS1 based normalised intensity values obtained from MaxQuant and is compatible with label free or TMT-tagged data (65). While spectral counts can be used for inferring protein abundance, we only focussed on the MS1 based methods largely to take advan-

tage of parametric statistical testing. As biological data is log normal for the most part, a log transformation can be applied to obtain normally distributed data which can be verified with histograms or quantile of quantile plots within the application. Common contaminants are filtered out from the dataset and groups are designated for further filtering. A few options are available such as filtering for valid values in one group, in each group or filtering all values. The latter is particularly useful for TMT-tagged data where the presence of a tag guarantees detection and very few values are missing. This becomes more nuanced in label-free proteomics experiments where it may be sensible to follow one of the other filtering approaches. As a mass spectrometer usually analyses the most abundant peptides in a given cycle the peptides corresponding to a protein may be seen in one instance but not another. Thus filtering for two occurrences of a protein out of three may be a valid approach to retain biologically relevant data. In an experiment such as treatment compared to control, the lack of a protein across all replicates in the control and the presence of the same protein in the treatment is likely of importance. In this case it is sensible to filter for valid values in one group and not each group. This does however cause a missing value problem which will have implications on hypothesis testing and corrections. To combat this ProVision uses a truncated normal distribution to impute the missing values, a method popularised by Perseus (62). The reasoning is that missing proteins are not missing at random, which is the assumption for other popular imputation methods such as K-nearest neighbours. Expression values for proteins in the filtered dataset are either missing because they were not expressed or expressed at a very low level compared to the other condition. Therefore, by estimating the lower quantile of the normal distribution associated with that sample a low expression value can be assigned (chapter 7, addendum A). After this step a hypothesis test can be conducted, ProVision uses LIMMA to run the statistical tests as this has been specifically developed for expression type data (66). Statistical tests without corrections are not allowed within the platform. The reasoning behind this was to avoid type I errors (false positives) at all costs, at the expense of type II errors (false negatives). The correction is however set to the most lenient to combat the number of type II errors that can occur. While a balance between type I and type II errors is an important consideration to any study, for the majority of cases one may want to avoid false positives as much as possible. This does however mean that critical information regarding the ground truth of an experiment may be lost in the process, especially if there is little difference between experiments. The data from the statistical tests is passed to the figure generator where ProVision provides volcano plots and heatmaps for direct visualisation of the differentially regulated proteins. This is further complemented by the use of the Webgestalt and STRING application programming interfaces to perform enrichments from the statistical tests as well as protein-protein interaction networks respectively (67,68). Detailed explanations of each analysis step

as well as each graph is recorded in tutorials located within the application to assist users at any step.

While ProVision is powerful in its own right, it still lacks key aspects that requires incorporation to be competitive. These include support for SILAC based applications as well as analysis for protein-protein interactions. Additional developments incorporating more advanced imputation techniques, especially in machine learning, would also greatly benefit the application. We released this application as an open source project for active development by community members who would like to assist in its growth. At the time of writing, ProVision has been accepted by members of the proteomics community and continues to grow, with the largest user base existing in Germany, USA and the Netherlands.

## FUTURE PERSPECTIVE AND CONCLUDING REMARKS

The work presented in this thesis has addressed questions in *M. tuberculosis* physiology, evolution, host-pathogen interactions and provided new tools to the broader community. There is however many unresolved questions that still remain as well as scope for improvement on current studies presented here. The current adoption of PacBio long read sequencing for *M. tuberculosis* genomes has the potential to add immense value to *M. tuberculosis* genomics. The widescale adoption of this technology will create robust genomes which will likely add to the current understanding of *M. tuberculosis* evolution. Long read sequencing technology also has the potential to generate lineage specific reference sequences to form the basis of other high throughput studies such as transcriptomics, proteomics among others. The discovery of gene fusions was done using a sub-optimal technique, albeit one that works for this specific chimera formation in *M. tuberculosis*. With the combination of long read instead of short read sequence data, many more genetic features can be detected including but not limited to the gene fusions. It is likely that there are other structural variations that create novel genetic features which *M. tuberculosis* is able to utilise. The genome is the basis of all functionality within any organism. By increasing efforts to investigate these hidden features within the genome may uncover factors with profound effects across other systems within the pathogen.

Proteomics is undoubtedly a powerful tool for the study of system wide changes to an organism in response to stimuli. We attempted SILAC based proteomics in auxotrophic *M. tuberculosis* and while we were able to label the proteome of *M. tuberculosis,* the detection of peptides were limited by the use of heavy leucine. While this strain

provides a large impact on the research capabilities of *M. tuberculosis,* especially in the proteomics space, it is still limited. Trypsin is commonly used to cut peptides and the recognition sites are either at lysine or arginine. This guarantees one of these amino acids are located on the peptide when sequenced and can be accurately quantified. Thus if a peptide does not contain leucine, it will not be sequenced and does not contribute to the quantification. Future studies in *M. tuberculosis* proteomics would benefit by creating a lysine/arginine double auxotroph. This may negate the safety of this organism outside a biosafety level three facility which would require additional investigation. However, if successful the creation of such a strain has the potential to become a mainstay in *M. tuberculosis* proteomics and other stable isotope based research.

We also investigated the proteome response of *M. tuberculosis ppe38-ppe71* mutants both *in vitro* as well as the macrophage temporal response to this strain (see chapter 5). It was known that PPE38 is associated with the secretion of a number of PE-PGRS proteins, but the exact state of the bacteria was unknown. We could show the presence of specific PE/PPE proteins, including PPE-MPTR proteins that were clearly controlled by PPE38, however the cell membrane also demonstrated a large number of intracellular proteins differentially regulated. This is likely caused by instability of the cell membrane in the absence of PE-PGRS and PPE-MPTR proteins on the surface. In addition, while PPE38 is clearly a chaperone for the PE-PGRS and PPE-MPTR proteins the exact mechanism is still unknown. There was no differential regulation in the cytoplasm of *M. tuberculosis* wild type and mutant strains, indicating that the secretion mechanism likely occurs within the membrane. The compromised cell wall and mechanism of secretion still remains unresolved. Understanding how the PE/PPE proteins interact and influence the mycomembrane and/or capsule could yield interesting observations regarding the fluidity and integrity of this structure.



With regards to the innate immune response to *ppe38-ppe71* deficient mycobacteria, we demonstrated the signalling pathway involved and could show that the inflammation is mediated by a p50/RelB pathway. This indicates a possible molecular switching mechanism between inflammatory states, which has been demonstrated before using other intracellular pathogens (51). It may be of interest to investigate the accessibility of this switching mechanism to wild type *M. tuberculosis.* If PPE38 expression is modulated during the course of infection this may indicate that the pathogen has the ability to influence macrophage polarisation indirectly. By further investigating the role of PPE38 in the innate immune response, especially in the context of polarisation states has potential in understanding not only how immune responses can be modulated but also how this pertains to bacterial survival *in vivo*. We however did not definitively show that a switch to M2 polarisation state occurs when infected with *M. tuberculosis*

*ppe38-ppe71* mutants. It would be important to understand whether inflammation is reduced or a full switch occurs, which could link the PE/PPE proteins with persistence *in vivo*.

In conclusion, the adoption of high throughput technologies and establishing standardised workflows is ushering a new era of biological research. When used correctly, these technologies provide insight into cellular dynamics that was not previously possible. By using these techniques we have the potential to investigate major cellular systems directly. Here we found that even simple bacterial organisms has a high degree of complexity. Often as molecular biologists we focus on a single protein, decipher its functionality and infer it on the greater system. There is however a homeostasis within the cell with many inter-dependent processes. It is thus highly likely that alterations in this homeostasis, be it genetically or phenotypically can reverberate throughout the cellular processes. While we have come a long way in understanding the physiology of *M. tuberculosis* there is still a long way to go. With the increased availability of data and data generating techniques we are now more than ever equipped to tackle major problems such as infectious disease. In the case of *M. tuberculosis*, it still has a presence among the top causes of death from an infectious agent and has held the title since antiquity. With technological advances we are increasing the resolution at which we are able to interrogate this pathogen. New techniques become available and new insights can be gained. By understanding fundamental processes as well as entire systems in context of infection we slowly build our understanding of *M. tuberculosis*. With the work presented in this thesis we have explored some of these processes, built upon others and created tools to assist future researchers in the fight against tuberculosis. With continued research and innovation we take strides in effectively combatting the spread and mortality of this pathogen.

## REFERENCES


1. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing [Internet]. Vol. 20, Genome Research. Cold Spring Harbor Laboratory Press; 2010 [cited 2020 Oct 30]. p. 1165–73. Available from: http://www.genome.org/cgi/doi/10.1101/gr.101360.109.
2. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci U S A [Internet]. 2016 Dec 27 [cited 2020 Oct 30];113(52):E8396–405. Available from: /pmc/articles/PMC5206522/?report=abstract
3. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol [Internet]. 2002;184. Available from: http://dx.doi.org/10.1128/JB.184.19.5479-5490.2002
4. Alland D, Lacher DW, Hazbón MH, Motiwala AS, Qi W, Fleischmann RD, et al. Role of Large Sequence Polymorphisms (LSPs) in Generating Genomic Diversity among Clinical Isolates of Mycobacterium tuberculosis and the Utility of LSPs in Phylogenetic Analysis. J Clin Microbiol [Internet]. 2007 Jan 1 [cited 2018 Mar 29];45(1):39–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17079498
5. Harper JW, Bennett EJ. Proteome complexity and the forces that drive proteome imbalance [Internet]. Vol. 537, Nature. Nature Publishing Group; 2016 [cited 2020 Oct 30]. p. 328–38. Available from: https://www.nature.com/articles/nature19947
6. Mann M, Kulak NA, Nagaraj N, Cox J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes [Internet]. Vol. 49, Molecular Cell. Elsevier; 2013 [cited 2020 Oct 30]. p. 583–90. Available from: http://dx.doi.org/10.1016/j.molcel.2013.01.029
7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol [Internet]. 2012 May 1 [cited 2020 Oct 30];19(5):455–77. Available from: /pmc/articles/PMC3342519/?report=abstract
8. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics [Internet]. 2014 Jun 15 [cited 2020 Oct 30];30(12):1660–6. Available from: https://academic.oup.com/bioinformatics/article/30/12/1660/380938
9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol [Internet]. 2011 Jul [cited 2020 Oct 30];29(7):644–52. Available from: /pmc/articles/PMC3571712/?report=abstract
10. Vyatkina K. De novo sequencing of top-down tandem mass spectra: A next step towards retrieving a complete protein sequence. Proteomes [Internet]. 2017 Mar 1 [cited 2020 Oct 30];5(1). Available from: /pmc/articles/PMC5372227/?report=abstract
11. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies [Internet]. Vol. 11, Nature Methods. Nature Publishing Group; 2014 [cited 2020 Oct 30]. p. 1114–25. Available from: https://www.nature.com/articles/nmeth.3144
12. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJL, Tabb DL. TagRecon: High-throughput mutation identification through sequence tagging. J Proteome Res [Internet]. 2010 Apr 5 [cited 2020 Oct 30];9(4):1716–26. Available from: https://pubs.acs.org/doi/abs/10.1021/pr900850m

13. Ma B, Johnson R. De novo sequencing and homology searching [Internet]. Vol. 11, Molecular and Cellular Proteomics. American Society for Biochemistry and Molecular Biology; 2012 [cited 2020 Oct 30]. Available from: http://www.mcponline.org
14. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature [Internet]. 2014 Sep 18 [cited 2020 Oct 30];513(7518):382–7. Available from: https://www.nature.com/articles/nature13438
15. Heunis T, Dippenaar A, Warren RM, van Helden PD, van der Merwe RG, Gey van Pittius NC, et al. Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium tuberculosis* Isolates. J Proteome Res [Internet]. 2017 Oct 6 [cited 2018 Feb 27];16(10):3841–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28820946
16. Stinear TP, Seemann T, Harrison PF, Jenkin GA, Davies JK, Johnson PD, et al. Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. Genome Res [Internet]. 2008;18. Available from: http://dx.doi.org/10.1101/gr.075069.107
17. Namouchi A, Didelot X, Schock U, Gicquel B, Rocha EPC. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res [Internet]. 2012 Apr 1 [cited 2018 Mar 16];22(4):721–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22377718
18. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A [Internet]. 1997 Sep 2 [cited 2017 Oct 12];94(18):9869–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9275218
19. Brosch R, Gordon S V, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2002 Mar;99(6):3684–9.
20. San Millan A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context [Internet]. Vol. 26, Trends in Microbiology. Elsevier Ltd; 2018 [cited 2020 Oct 30]. p. 978–85. Available from: https://doi.org/10.1016/j.tim.2018.06.007
21. Stucki D, Gagneux S. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. Tuberculosis [Internet]. 2013 Jan 1 [cited 2019 Sep 15];93(1):30–9. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1472979212002028
22. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. Vol. 26, Seminars in Immunology. Academic Press; 2014. p. 431–44.
23. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. J Bacteriol [Internet]. 1996 [cited 2020 Oct 30];178(5):1274–82. Available from: https://pubmed.ncbi.nlm.nih.gov/8631702/
24. Yang C, Luo T, Sun G, Qiao K, Sun G, Deriemer K, et al. Mycobacterium tuberculosis Beijing strains favor transmission but not drug resistance in China. Clin Infect Dis [Internet]. 2012 Nov 1 [cited 2020 Oct 30];55(9):1179–87. Available from: https://academic.oup.com/cid/article/55/9/1179/436433
25. Tram TTB, Nhung HN, Vijay S, Hai HT, Thu DDA, Ha VTN, et al. Virulence of Mycobacterium tuberculosis Clinical Isolates Is Associated With Sputum Pre-treatment Bacterial Load, Lineage, Survival in Macrophages, and Cytokine Response. Front Cell

Infect Microbiol [Internet]. 2018 [cited 2020 Oct 30];8:417. Available from: /pmc/articles/PMC6277702/?report=abstract

26. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nat Microbiol [Internet]. 2018 Feb 15 [cited 2018 Mar 6];3(2):181–8. Available from: http://www.nature.com/articles/s41564-017-0090-6
27. Simeone R, Sayes F, Song O, Gröschel MI, Brodin P, Brosch R, et al. Cytosolic Access of Mycobacterium tuberculosis: Critical Impact of Phagosomal Acidification Control and Demonstration of Occurrence In Vivo. Salgame P, editor. PLOS Pathog [Internet]. 2015 Feb 6 [cited 2020 Nov 24];11(2):e1004650. Available from: https://dx.plos.org/10.1371/journal.ppat.1004650
28. Gao LY, Guo S, McLaughlin B, Morisaki H, Engel JN, Brown EJ. A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion. Mol Microbiol [Internet]. 2004 Sep 1 [cited 2020 Oct 30];53(6):1677–93. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2958.2004.04261.x
29. Armstrong JA. RESPONSE OF CULTURED MACROPHAGES TO MYCOBACTERIUM TUBERCULOSIS, WITH OBSERVATIONS ON FUSION OF LYSOSOMES WITH PHAGOSOMES. J Exp Med [Internet]. 1971 Sep 1 [cited 2016 Mar 9];134(3):713–40. Available from: http://jem.rupress.org/cgi/content/long/134/3/713
30. Sturgill-Koszycki S, Schlesinger PH, Chakraborty P, Haddix PL, Collins HL, Fok AK, et al. Lack of acidification in Mycobacterium phagosomes produced by exclusion of the vesicular proton-ATPase. Science (80- ) [Internet]. 1994 Feb 4;263(5147):678–81. Available from: http://science.sciencemag.org/content/263/5147/678.abstract
31. MacMicking JD, Taylor GA, McKinney JD. Immune Control of Tuberculosis by IFN-γ-inducible LRG-47. Science (80- ) [Internet]. 2003 Oct 24 [cited 2020 Oct 30];302(5645):654–9. Available from: https://pubmed.ncbi.nlm.nih.gov/14576437/
32. Manzanillo PS, Shiloh MU, Portnoy DA, Cox JS. Mycobacterium tuberculosis activates the DNA-dependent cytosolic surveillance pathway within macrophages. Cell Host Microbe [Internet]. 2012 May 17 [cited 2019 Mar 25];11(5):469–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22607800
33. Peng X, Sun J. Mechanism of ESAT-6 membrane interaction and its roles in pathogenesis of Mycobacterium tuberculosis. Toxicon [Internet]. 2016 Jun 15 [cited 2020 Oct 30];116:29–34. Available from: /pmc/articles/PMC4973572/?report=abstract
34. Houben ENGG, Korotkov K V., Bitter W. Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta [Internet]. 2014 Aug;1843(8):1707–16. Available from: http://dx.doi.org/10.1016/j.bbamcr.2013.11.003
35. Ates LS. New insights into the mycobacterial PE and PPE proteins provide a framework for future research [Internet]. Vol. 113, Molecular Microbiology. Blackwell Publishing Ltd; 2020 [cited 2020 Oct 30]. p. 4–21. Available from: /pmc/articles/PMC7028111/?report=abstract
36. Abdallah AM, Verboom T, Weerdenburg EM, Gey van Pittius NC, Mahasha PW, Jiménez C, et al. PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. Mol Microbiol [Internet]. 2009 Aug 1 [cited 2019 Apr 18];73(3):329–40. Available from: http://doi.wiley.com/10.1111/j.1365-2958.2009.06783.x
37. Chaitra MG, Shaila MS, Nayak R. Detection of interferon gamma-secreting CD8+ T lymphocytes in humans specific for three PE/PPE proteins of Mycobacterium tuberculosis.

Microbes Infect [Internet]. 2008 Jul [cited 2020 Oct 30];10(8):858–67. Available from: https://pubmed.ncbi.nlm.nih.gov/18653370/

38. Chaitra MG, Shaila MS, Nayak R. Evaluation of T-cell responses to peptides with MHC class I-binding motifs derived from PE_PGRS 33 protein of Mycobacterium tuberculosis. J Med Microbiol [Internet]. 2007;56. Available from: http://dx.doi.org/10.1099/jmm.0.46928-0
39. Sampson SL. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol [Internet]. 2011;2011(Figure 1):497203. Available from: http://dx.doi.org/10.1155/2011/497203
40. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jage BB. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393(6685):537–544.
41. McEvoy CRE, van Helden PD, Warren RM, van Pittius N, Gey van Pittius NC. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. BMC Evol Biol [Internet]. 2009 Jan;9(1):237. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-237
42. Hanekom M, Gey Van Pittius NC, McEvoy C, Victor TC, Van Helden PD, Warren RM. Mycobacterium tuberculosis Beijing genotype: A template for success [Internet]. Vol. 91, Tuberculosis. Tuberculosis (Edinb); 2011 [cited 2020 Oct 30]. p. 510–23. Available from: https://pubmed.ncbi.nlm.nih.gov/21835699/
43. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet [Internet]. 2013 Jul [cited 2020 Oct 30];45(7):784–90. Available from: https://pubmed.ncbi.nlm.nih.gov/23749189/
44. Gengenbacher M, Kaufmann SHE. Mycobacterium tuberculosis : Success through dormancy. FEMS Microbiol Rev. 2013;36(3):514–32.
45. Dong D, Wang D, Li M, Wang H, Yu J, Wang C, et al. PPE38 modulates the innate immune response and is required for Mycobacterium marinum virulence. Infect Immun [Internet]. 2012 Jan [cited 2019 Mar 7];80(1):43–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22038915
46. Ates LS, Sayes F, Frigui W, Ummels R, Damen MPM, Bottai D, et al. RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. PLoS Pathog [Internet]. 2018 [cited 2019 Mar 7];14(6):e1007139. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29912964
47. Baker RG, Hayden MS, Ghosh S. NF-κB, inflammation, and metabolic disease [Internet]. Vol. 13, Cell Metabolism. NIH Public Access; 2011 [cited 2020 Oct 30]. p. 11–22. Available from: /pmc/articles/PMC3040418/?report=abstract
48. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther [Internet]. 2017 Jul 14 [cited 2019 Mar 29];2:17023. Available from: http://www.nature.com/articles/sigtrans201723
49. Turner MD, Nedjai B, Hurst T, Pennington DJ. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. Vol. 1843, Biochimica et Biophysica Acta - Molecular Cell Research. Elsevier; 2014. p. 2563–82.

50. Sun S-C. The non-canonical NF-κB pathway in immunity and inflammation. Nat Rev Immunol [Internet]. 2017 Jun 5 [cited 2019 Mar 26];17(9):545–58. Available from: http://www.nature.com/doifinder/10.1038/nri.2017.52
51. Stapels DAC, Hill PWS, Westermann AJ, Fisher RA, Thurston TL, Saliba AE, et al. Salmonella persisters undermine host immune defenses during antibiotic treatment. Science (80- ). 2018 Dec 7;362(6419):1156–60.
52. Burggraaf MJ, Speer A, Meijers AS, Ummels R, Van Der Sar AM, Korotkov K V., et al. Type VII secretion substrates of pathogenic mycobacteria are processed by a surface protease. MBio. 2019 Oct 29;10(5).
53. The Aetiology of Tuberculosis | American Review of Tuberculosis [Internet]. [cited 2020 Oct 30]. Available from: https://www.atsjournals.org/doi/abs/10.1164/art.1932.25.3.285?journalCode=art
54. L. Ramakrishnan, N. A. Federspiel and SF, Ramakrishnan L, Federspiel NA, Falkow S. Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. Science. 2000 May;288(5470):1436–1439.
55. Grover S, Sharma T, Singh Y, Kohli S, Manjunath P, Singh A, et al. The PGRS domain of Mycobacterium tuberculosis PE_PGRS protein Rv0297 is involved in Endoplasmic reticulum stress-mediated apoptosis through toll-like receptor 4. MBio [Internet]. 2018 May 1 [cited 2020 Jul 10];9(3). Available from: /pmc/articles/PMC6016250/?report=abstract
56. Chihota VN, Müller B, Mlambo CK, Pillay M, Tait M, Streicher EM, et al. Population structure of multi- and extensively drug-resistant Mycobacterium tuberculosis strains in South Africa. J Clin Microbiol [Internet]. 2012 Mar [cited 2020 Oct 30];50(3):995–1002. Available from: /pmc/articles/PMC3295122/?report=abstract
57. Naidoo P, Theron G, Rangaka MX, Chihota VN, Vaughan L, Brey ZO, et al. The South African Tuberculosis Care Cascade: Estimated Losses and Methodological Challenges. In: Journal of Infectious Diseases [Internet]. Oxford University Press; 2017 [cited 2020 Oct 30]. p. S702–13. Available from: /pmc/articles/PMC5853316/?report=abstract
58. Luzze H, Johnson DF, Dickman K, Mayanja-Kizza H, Okwera A, Eisenach K, et al. Relapse more common than reinfection in recurrent tuberculosis 1-2 years post treatment in urban Uganda. Int J Tuberc Lung Dis [Internet]. 2013 Mar 1 [cited 2020 Oct 30];17(3):361–7. Available from: /pmc/articles/PMC6623981/?report=abstract
59. Gillespie SH, Crook AM, McHugh TD, Mendel CM, Meredith SK, Murray SR, et al. Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. N Engl J Med [Internet]. 2014 Oct 23 [cited 2020 Oct 30];371(17):1577–87. Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa1407426
60. Moosazadeh M, Bahrampour A, Nasehi M, Khanjani N. The incidence of recurrence of tuberculosis and its related factors in smear-positive pulmonary tuberculosis patients in Iran: A retrospective cohort study. Lung India [Internet]. 2015 Nov 1 [cited 2020 Oct 30];32(6):557–60. Available from: /pmc/articles/PMC4663856/?report=abstract
61. Martin TR, Frevert CW. Innate immunity in the lungs. In: Proceedings of the American Thoracic Society [Internet]. American Thoracic Society; 2005 [cited 2020 Oct 30]. p. 403–11. Available from: /pmc/articles/PMC2713330/?report=abstract
62. Tyanova S, Cox J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. In: Methods in molecular biology (Clifton, NJ) [Internet]. 2018 [cited 2019 Oct 13]. p. 133–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29344888

63. Gatto L, Breckels LM, Naake T, Gibb S. Visualization of proteomics data using R and Bioconductor [Internet]. Wiley-VCH Verlag; Apr, 2015 p. 1375–89. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pmic.201400392
64. Gatto L, Lilley KS. Msnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics. 2012 Jan;28(2):288–9.
65. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008 Dec 30 [cited 2018 Feb 27];26(12):1367–72. Available from: http://www.nature.com/articles/nbt.1511
66. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Apr 20 [cited 2018 Mar 15];43(7):e47–e47. Available from: http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for
67. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res [Internet]. 2019 [cited 2020 May 26];47:199–205. Available from: https://academic.oup.com/nar/article-abstract/47/W1/W199/5494758
68. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res [Internet]. 2018 [cited 2020 May 26];47:607–13. Available from: https://string-db.org/.

# A

## Nederlandse samenvatting
## Afrikaanse opsomming
## Dankwoord
## Curriculum vitae
## Scholarships and grants
## List of publications

# NEDERLANDSE SAMENVATTING

## PORTRET VAN EEN PATHOGEEN: EEN DIEPGAANDE KARAKTERISERING VAN *M. TUBERCULOSIS* EN ZIJN GASTHEER MET BEHULP VAN MULTIDIMENSIONALE PROTEOMICS.

### Inleiding en doel van dit onderzoek.

*Mycobacterium tuberculosis* is de oorzaak van de infectieziekte tuberculose bij de mens en is een van de dodelijkste infectieziektes wereldwijd. In het jaar 2020 waren er 130 nieuwe tuberculose gevallen per 100.000 mensen en 1.2 miljoen doden (WHO tuberculosis report 2020). Ongeacht aanzienlijke inspanningen om de verspreiding van de ziekte te stoppen, is er nog steeds een wereldwijd probleem. Landen met een hoge incidentie van resistente bacteriën, zoals Zuid-Afrika, zijn kwetsbaar, dus is het belangrijk om verder onderzoek te doen.

*M. tuberculosis* is een gevaarlijke gespecialiseerde bacterie die voornamelijk longziekte veroorzaakt. Het heeft een zeer klein genoom dat speciaal is ontwikkeld om in een menselijke macrofaag, een afweercel, te leven. *M. tuberculosis* wordt vaak gevonden in het midden van meerdere immuuncellen, waarbij de cellen en bacterien samen een granuloom vormen. Het granuloom is een kenmerk van klinische tuberculose en meerdere granulomen worden gevonden in het infectiegebied. De gelijktijdige evolutie van de bacteriën en de menselijke cellen zorgt voor een hoge specialisatie in beide. Die bacteriële specialisatie is terug te zien in het kleine genoom en de eiwitten. De eiwitten zijn verantwoordelijk voor zowel de structuur als de metabole functie van de bacterie. Door de bijzondere evolutie van *M. tuberculosis* zijn de unieke kenmerken van het genoom en het proteoom zeer belangrijk voor het onderscheid tussen mycobacteriën die wel en geen ziekte veroorzaken. Dit verschil kan ons aanwijzingen geven voor de juiste behandeling van de ziekte tuberculose. Het type VII secretiesysteem, met name het ESX-5 systeem, is een uniek kenmerk van de ziekte veroorzakende mycobacteriën. Dit systeem is essentieel voor de groei van *M. tuberculosis* en is verantwoordelijk voor de secretie van unieke eiwitten, namelijk de PE-PGRS en PPE-MPTR eiwitten. Deze eiwitten worden in grote aantallen buiten de cel aangetroffen, zijn uniek voor langzaam groeiende mycobacteriën en hebben het vermogen om het immuunsysteem van de gastheer te beïnvloeden. Er is zeer weinig bekend over de samenwerking van de PE-PGRS en PPE-MPTR als een eenheid en of er werkelijk een samenwerking is. In dit proefschrift wordt een mutant van het *ppe3*8 gen bestudeerd. In een eerdere studie is gevonden dat het eiwit PPE38 de secretie van de PE-PGRS eiwitten uitschakelt. Een

mutant van *ppe38,* waarbij dit eiwit dus niet meer functioneert, heeft een groot effect op de interactie tussen *M. tuberculose* en zijn gastheer. Het is bekend dat een verlies van PPE38 zorgt voor meer virulente of agressieve *M. tuberculosis* en deze vorm komt voor in de natuur. De effecten van deze mutant op het cellulaire vlak zijn een onderdeel van het werk beschreven in dit proefschrift.

Het doel van het onderzoek beschreven in dit proefschrift is om meer inzicht te krijgen in de systeem biologie van zowel *M. tuberculosis* als zijn gastheer. Hiermee kunnen diverse hulpmiddelen en technieken voor verder onderzoek worden opgezet en toegepast. In dit proefschrift wordt gebruik gemaakt van een auxotrofe stam van *M. tuberculosis* voor onder andere geavanceerde massaspectrometrie, inzicht in de evolutiebiologie van *M. tuberculosis,* effecten van *M. tuberculosis* op de macrofaag en directe observaties van de effecten van de infectie op longweefsel. Tot slot hebben we de analyse van proteomics data verbeterd door een online platform te creëren.

## Gebruikte onderzoeksmethoden.

In dit proefschrift gebruiken we in ieder hoofdstuk proteomics in zowel *M. tuberculosis* als zijn gastheer. In hoofdstukken 2 en 3 gebruiken we een auxotrofe variant van *M. tuberculosis,* welke een specifieke toevoeging nodig heeft om te groeien, waardoor het veel veiliger is om de stam in een ML-2 laboratorium te gebruiken voor complexe proteomics experimenten. Deze hoofstukken zijn essentieel voor de opzet van proteomics technieken die worden gebruikt in andere hoofstukken. Onder andere wordt in deze hoofdstukken laten zien dat het van belang is de bacteriën op een specifieke manier te laten groeien, namelijk zonder abumine of polymeren.

Naast de auxotrofe *M. tuberculosis*, gebruiken we natuurlijke stammen van *M. tuberculosis* die zijn geïsoleerd uit een geïnfecteerd individu. Voor gecontroleerde experimenten waar de genetische achtergrond belangrijk is, gebruiken we de *M. tuberculosis* H37Rv of CDC1551 laboratoriumstammen of een *ppe38* knock-out mutant. In hoofdstuk 5 worden laboratoriumvarianten van menselijke macrofagen gebruikt en in hoofdstuk 6 maken we gebruik van patiëntmateriaal. Het patiëntmateriaal is afkomstig van alveolaire bronchiale spoelsels uit de longen van mensen tijdens *M. tuberculosis* besmetting en na behandeling.

Proteomics wordt vaak gebruik voor het bepalen van het verschil in hoeveelheid eiwit tussen verschillende condities. Er zijn echter veel meer toepassingen en manieren om de nauwkeurigheid van een proteomics experiment te verbeteren. In dit proefschrift gebruiken we labelvrije proteomics voor het merendeel van de studies, evenals zogenaamde “tandem mass tagged” (TMT) en “stable isotopes of aminoacids in cell culture”

(SILAC) gebaseerde proteomics. Met deze technieken in combinatie met goede data analyse en statistiek is het mogelijk om betrouwbare data te verwerven. Daarnaast kunnen deze technieken worden toegepast op elk organisme en wordt het daarom vaak gebruikt, onder andere in bijna elk hoofdstuk van dit proefschrift. Het gebruik van SILAC en TMT-labels wordt toegepast in de hoofdstukken 3, 5 en 6. Met SILAC is het mogelijk om meer complexe experimenten met hoge nauwkeurigheid uit te voeren. We gebruiken SILAC om post-translationale modificaties te bepalen in hoofdstuk 3 en om eiwitvernieuwing te bepalen in zowel hoofdstuk 3 als 5.

## Resultaten in dit proefschrift.

In **hoofdstuk 2** wordt de auxotrofe mutant van *M. tuberculosis* gekarakteriseerd in vergelijking met de *M. tuberculosis* H37Rv laboratorium stam. De boodschap van dit hoofdstuk is dat de auxotrofe stam veilig is en als model voor *M. tuberculosis* kan worden gebruikt. Het is wel zo dat levende organismen een hoge mate van complexiteit hebben, dus het is mogelijk dat zelfs een kleine verandering op het genoom vlak een groot effect heeft. Daarom is het essentieel om te bepalen of er effecten zijn van de auxotrofe stam ten opzichte van de normale *M. tuberculosis* stam. We bestuderen de verschillen op het genoom vlak, verschillen in groei zowel *in vitro* als *ex vivo*, effecten van antibiotica, proteoomreactie op zuur stress en macrofaagreactie op infectie met de auxotrofe stam. Dit hoofdstuk laat zien dat er weinig verschillen zijn, behalve welke al werden verwacht. Het grootste verschil is de reacties op zuurstress, de auxotroof is iets gevoeliger voor deze stress wat leidt tot een matig effect op groei.

**Hoofdstuk 3** is het vervolg op hoofdstuk 2. De auxotrofe stam is een leucine auxotroof en heeft daarom geen vermogen om zelf leucine te produceren. De enige optie is om het leucine vanuit de omgeving op te nemen. Het is daarom mogelijk om de leucine in het groeimedium te vervangen met leucine dat atomair zwaarder is. Het zware leucine is een stabiele variant en kan met gemak worden teruggevonden in een massaspectrometer. We kunnen SILAC gebruiken om meer inzicht te krijgen in de auxotrofe stam en zijn reactie op zuur stress. Naast verschillen in hoeveelheid in verloop van tijd, hebben we ook de post-translationele modificaties en eiwitomzet bestudeert. We vinden dat na verloop van tijd een groot deel van de eiwit concentraties verlaagd zijn, behalve enkele eiwitten die gerelateerd zijn aan pathogenese, o.a die type VII secretie systeem. We vonden ook dat de type VII secretie een belangrijke rol speelt in de post-translationale modificaties, met name bij fosforylering om die ESX-1 componenten en effectoren. Tot slot was de eiwitomzet een grote verrassing, er was namelijk tijdens stress bijna geen opname van leucine. Daarom konden we niet bepalen wat het verschil in omzet was tussen de stressreacties. Het is interessant dat dit gebeurt en het betekent dat de bacteriën snel iets belangrijks zoals aminozuur opname afsluiten. Wij kunnen niet verklaren

waarom dit proces zo snel gestopt wordt. Het was wel duidelijk dat tijdens meerdere stress reacties het secretie systemen een grote rol had in het beïnvloeden hiervan.

In **hoofdstuk 4** gebruiken we genoom data van klinische *M. tuberculosis* stammen die afkomstig zijn van patiënten uit die West-Kaapse regio van Zuid-Afrika. Allereerst hebben we bestudeerd hoe vaak een mutatie in de *ppe38* regio in verschillende geslachtslijnen van *M. tuberculosis* voorkomt. Uit deze experimenten ontdekten we dat de meeste *ppe38* knock-out mutant voorkomen in lijn 2 en in een mindere mate in lijn 4. Met nadere inspectie van de DNA breekpunten vonden wij een verschil in de DNA volgorde tussen de twee lijnen. Zo bleek dat in een aantal gevallen lijn 4 een nieuw *ppe38* gen maakt die functioneel is, wat zorgt voor het maken van een 'fusie-eiwit'. We hebben dit opgevolgd door een software programma te schijven dat dit soort breekpunten in de DNA-volgorde kon identificeren. Met deze informatie konden we meer breuken vinden die leiden tot een hervorming van genen in het genoom. Dit werd vervolgens ook gebruikt om een proteomics-database op te zetten voor het vinden van meer fusie-eiwitten. Een opmerkelijk voorbeeld was de hervorming van *Rv2623-Rv2628* in lijn 2 wat mogelijk een invloed op groei hebben. We konden hier dus bewijzen dat *M. tuberculosis* bij afwezigheid van horizontale gen-overdracht nog steeds nieuw genetisch materiaal kan aanmaken door gebruik te maken van bestaand materiaal.

In **hoofdstuk 5** wordt onderzoek gedaan naar de *ppe38* mutant en het effect ervan op het aangeboren immuunsysteem, specifiek met betrekking tot de macrofaag. Met behulp van massaspectrometrie zagen we dat de macrofaag een veel lagere response vertoont tijdens infectie met de *ppe38* mutant. Verder wordt aangetoond dat deze reacties optreden als gevolg van een dysbalans in het Nuclear Factor kappa B (NK-kB) eiwit. Met behulp van confocale microscopie hebben we het signalering pad bestudeerd en het verschil in translocatie van Rel eiwitten naar de celkern kunnen vinden. Het levert interessante resultaten op over de mogelijke invloed van deze mutant op de respons van het immuunsysteem. Vervolgwerk met ELISA tests bevestigde inderdaad dat er factoren worden geproduceerd en uitgescheiden door cellen geinfecteerd met de *ppe38* mutant welke inflammatie kunnen afremmen. Het is daarom mogelijk dat een verlies in secretie van de PE-PGRS eiwitten leidt tot een lagere initiële immuunrespons die na verloop van tijd bijdraagt aan een groter aantal geinfecteerde cellen bij geïnfecteerde individuen. Dit is een kenmerk is van infectie met lijn 2 mycobacterien.

**Hoofdstuk 6** beschrijft een eiwitprofiel in longweefsel van patiënten met actieve tuberculose en van patiënten na afloop van behandeling. We gebruiken een aantal metingen, o.a. het aantal bacteriën, FDG PET-CT scans en alveolaire bronchiale spoelsel/lavage (BAL) van de patiënt. We vinden hier dat de meerderheid van patiënten na de

behandeling geen aantoonbare bacteriën meer heeft. Er waren echter ook patiënten met aantoonbare bacteriën alleen in de BAL maar niet in hun speeksel. Deze patiënten worden op basis van de uitslag van de speekselkweektest als gezond beschouwd, maar het is met tekenen van bacteriën in de long en op FDG PET-CT een ander verhaal. Daarnaast konden we voorspellen met proteomics dat een van de patiënten enkele maanden na het bezoek aan het ziekenhuis opnieuw ziek zou worden van de mycobacteriele besmetting. Het is daarom mogelijk dat patiënten zich gezond presenteren maar dat dit niet de juiste conclusie is. Dit is belangrijk omdat het potentiaal dodelijk is, al is het in een klein deel van de gevallen. Met proteomics is het ook mogelijk om te ontcijferen welke patiënten een risico op heractivatie van de ziekte hebben. Het gebruik van proteomics voor dit doel is op dit moment helaas niet mogelijk in de dagelijkse praktijk vanwege de complexiteit van deze techniek, maar met verder onderzoek in het veld kan dit op andere manieren mogelijk wel.

**Hoofdstuk 7** is de afronding van alle proteomics-kennis die in dit proefschrift wordt gebruikt. Hoewel veel biologische vragen kunnen worden beantwoord, hangt het allemaal af van de data analyse erachter. Proteomics genereert een grote hoeveelheid data, waardoor het bijna onmogelijk is om op meer traditionele manieren te analyseren. De analyse hangt daarnaast ook af van het soort experimenten dat wordt uitgevoerd. Programmeren is nodig om goed inzicht te krijgen in de onderliggende data. Dit levert voor veel onderzoekers een probleem op, omdat het programmeren en geavanceerde statistiek bijna geen deel uitmaken van standaard opleiding. Om dit te verbeteren en te delen met collega's in het veld hebben we alles wat we op dit gebied hebben geleerd gebundeld in een web gebaseerd data-analyseplatform. Dit hoofdstuk en het bijbehorende addendum zijn de beschrijvingen van dit platform. Vanaf de releasedatum wordt het platform veel gebruikt door meerdere mensen met het grootste gebruikersbestand in de VS, Duitsland, Nederland en het VK.



## Conclusie

Ten slotte luiden de aanpassing van technologieën zoals proteomics en genomics, en de standaardisatie van werk, een nieuw tijdperk in voor biologisch onderzoek. Met het juiste gebruik van deze technologie kunnen nieuwe inzichten worden opgedaan in cellulaire dynamiek die voorheen niet mogelijk waren. De technologie maakt het mogelijk om grootschalige systemen direct te bestuderen. Met deze technologie hebben we ontdekt dat zelfs eenvoudige organismen zoals bacteriën een hoog niveau van complexiteit hebben waarin veel geheimen verborgen liggen. Wat betreft de fysiologie en evolutionaire kenmerken van *M. tuberculosis en* zijn varianten, moeten we nog veel leren. In dit proefschrift hebben we meer inzicht gekregen in belangrijke systemen van de mycobacteriën. Het is vooral duidelijk dat de bacterie moet worden bestudeerd in

de context van zijn omgeving of gastheer voor het beste inzicht in zijn gedrag. Met de combinatie van systeembrede karakterisering en begrip van de fundamentele processen waardoor *M. tuberculosis* zich onderscheidt van andere pathogenen, kunnen we hopen effectief te zijn in de strijd tegen deze ziektewekker.

# AFRIKAANSE OPSOMMING

## PORTRET VAN 'N PATOGEEN: 'N KARAKTERISERING VAN *M. TUBERCULOSIS* ASOOK SY GASHEER MET BEHULP VAN MULTIDIMENSIONELE PROTEOMIKA

### Inleiding tot hierdie navorsing en die doel.

*Mycobacterium tuberculosis* is die bakteriese agent wat die aansteeklike siekte tuberkulose veroorsaak in mense. Tuberkulose is nog steeds een van die dodelikste siektes in die konteks van infeksie siektes. In 2020, was daar 130 nuwe tuberkulose gevalle per 100 000 mense en 1.2 miljoen mense is dood as 'n direk gevolg van tuberkulose. Ongeag noemenswaardige pogings om die verspreiding van die siekte te stop is daar wêreld wyd 'n probleem. Dit is veral opvallend in lande soos Suid Afrika, waar daar hoë voorkoms is van dwelm weerstandbare mikobacterieë, dus is dit van uiters belang om verder navorsing te doen.

*M. tuberculosis* is 'n gevaarlike gespesialiseerde bakterie wat hoofsaaklik long infeksies veroorsaak. Die bakterie het 'n klein genoom wat spesifiek ontwikkel het om te lewe in 'n mens makrofaag sel. *M. tuberculosis* is dikwels omsingel deur meerdere immuun selle wat 'n granuloom vorm. Inderdaad is die granuloom 'n spesifieke kliniese kenmerk van *M. tuberculosis* infeksie. Hierdie gelyktydige ontwikkeling van *M. tuberculosis* en mens selle lei tot 'n so genaamd wapen wedloop waar die bakterieë hoog gespesialiseerd word in beide sy genoom en uitdrukking van sy gene. In alle lewendige organismes vorm die proteïene die funksionele onderdele van die sel. Dus beide die struktuur van die sel en die metaboliese funksies word verrig deur sy proteïene. Omdat die mikobakterieë so 'n besondere evolusionêre ontwikkeling gehad het, is die studie van die unieke kenmerke van die genoom en die proteoom van belang. Veral kenmerke wat siekte veroorsakende mikrobakterie skei van die wat nie-patogenies is, is dus moontlik die sleutel tot 'n optimale antibiotika. Die tipe VII sekresie sisteem van *M. tuberculosis* is naamlik so 'n kenmerk. Hierdie stelsel is uniek tot die patogeniese mikobacterieë, veral die ESX-5 sisteem het onlangs ontwikkel in die evolusionêre tydlyn. Dié sisteem is verantwoordelik vir die sekresie van die PE/PPE proteïene, 'n groot aantal unieke proteïene wat byna 10 persent van die genoom uitdrukkings potensiaal is. Dit is voor die hand liggend en voorheen bewys dat hierdie proteïene die vermoë het om in te meng met die immuun prosesse van die gasheer selle. Die manier waarby dit gehandhaaf word op so 'n groot skaal is nog steeds onbekend. In hierdie tesis word veral 'n uitslaan mutant van *ppe38* in diepte bestudeer. Dit is juis omdat hierdie mutant die sekresie van

'n groot aantal ander PE-PGRS ('n sub-familie van pe/ppe) proteïene uitskakel. Met so 'n groot verlies is die *ppe38* mutant 'n aantreklike teiken vir studie, veral siende dat hierdie mutasie ook natuurlik voorkom.

Die doel van die navorsing wat hier verrig word is dus om uit te brei op die ope vrae wat in die sisteem biologie van beide *M. tuberculosis* asook sy gasheer van toepassing is. Deur so 'n studie is dit moontlik om beide die nodige gereedskap en tegnieke op te stel vir verdere navorsing asook duidelike riglyne van hoe om die tegnieke toe te pas. Dit het gelei tot die opstel van 'n auxotrofiese stam van *M. tuberculosis* wat gebruik kan word vir gevorderde massa spektrometrie, aanvullende insig in die evolusionêre biologie van *M. tuberculosis,* die biologie van *M. tuberculosis* in makrofage en die reaksie van die long tot infeksie. Om meerdere van die vrae te beantwoord het ons gebruik gemaak van proteomika, en daarom ook veel geleer in die veld, wat ons in staat gestel het om verder bydrae gelewer tot die breër gemeenskap deur 'n analitiese platform vry te stel.

## Navorsing metodes

Die navorsing in hierdie tesis draai om die sentrale toepassing van proteomika tot *M. tuberculosis* asook sy gasheer selle. In hoofstukke twee en drie gebruik ons 'n auxotrofiese variant wat veiliger is en in bioveiligheidsvlak 2 kondisies gebruik kan word. Dit laat ook toe vir meer komplekse eksperimente, veral in proteomika. Die studies vorm ook die fondasie vir die proteomika tegnieke wat toegepas word in ander hoofstukke. Onder andere is dit ook die spesifieke manier van bakteriese groei wat belangrik is, naamlik in media sonder albumien of polimere. Ons gebruik natuurlike stamme van *M. tuberculosis* wat vanuit geïnfekteerde individu geïsoleer was. Vir gekontroleerde eksperimente waar die genetiese agtergrond van belang is, gebruik ons die *M. tuberculosis* H37Rv of CDC1551 laboratorium stamme of 'n *ppe38* uitslaan mutant. Ons maak ook gebruik van laboratorium variante van mens makrofage in hoofstuk 5 asook pasiënt materiaal in hoofstuk 6. Die pasiënt materiaal is afkomstig van alveolêre brongiale spoel uit die longe van mense besmet met tuberkulose en na die afloop van behandeling.

Die mees algemene gebruik van proteomika is om die verskil in hoeveelheid proteïen te bepaal tussen toestande. Daar is wel veel meer toepassinge asook maniere om akkuraatheid van proteomika te verbeter. In hierdie tesis word ook gebruik gemaak van etiket vrye proteomika, so genaamd label-free proteomics in Engels, vir die meerderheid van studies, asook so genaamd "tandem mass tagged" (TMT) en "stable isotopes of amino acids in cell culture" (SILAC) gebaseerde proteomika. Al is etiket vrye proteomika minder akkuraat, is dit nogmaals moontlik om met data analise en eksperimentele praktyk goeie data te werf met hierdie tegniek. Die tegniek is toepaslik

op enige organisme en word daarom in amper al die hoofstukke van hierdie tesis gebruik, veral vir verskil in hoeveelheid bepaling. Die gebruik van stabiele aminosure en TMT etikette word in hoofstuk 3, 5 en 6 toegepas. Met stabiele swaar aminosure is dit verder moontlik om meer komplekse hoë akkurate eksperimente uit te voer. Ons maak gebruik van hierdie aminosure om in hoofstuk 3 na-translasionele modifikasies te bepaal en in beide hoofstuk 3 en 5 proteïen omset te bepaal.

## Resultate in die tesis

In **hoofstuk 2** word die auxotrofiese mutant van *M. tuberculosis* gekarakteriseer in vergelyking met die *M. tuberculosis* H37Rv wilde tipe laboratorium stam. Die auxotrofiese stam is veiliger dan die wilde tipe en kan gebruik word as 'n model vir die wilde tipe in bioveiligheidsvlak 2 laboratoriums. Siende dat lewende organismes 'n groot mate van kompleksiteit het, is dit moontlik dat matige veranderinge 'n groot effek kan hê. Daarom is dit noodsaaklik om te bepaal of daar enige newe effekte is op die auxotrofiese stam. Ons bestudeer die verskille in genoom, verskille in groei beide *in vitro* en *ex vivo*, effekte van teenmiddels, proteoom reaksie tot suur en die makrofaag reaksie tot die auxotrofiese stam. Ons vind dat daar nie soveel verskille is in die meerderheid van gevalle nie, behalwe wat al verwag was. Ons vind wel dat die auxotrofiese stam 'n groter verskil toon in sy response tot suur stres. Dié stam is meer gevoelig vir die stres response en daarom is daar matige verskil in sel verdeling ook.

**Hoofstuk 3** is die opvolg van hoofstuk twee. Die auxotrofiese stam is naamlik 'n leusien auxotroof en het nie die vermoë om leusien self te produseer nie. Dit is dus moontlik om die leusien in die groei media te vervang met leusien wat atomies swaarder is. Die swaar leusien is stabiel en kan maklik terug gevind word in massa spektrometrie. Ons het dus SILAC gebruik om meer insig te kry oor die auxotrofiese *M. tuberculosis* stam en sy response tot suur stres. Behalwe vir verskille in hoeveelheid oor tyd, het ons ook die na-translasionele modifikasies en proteïen omset. Ons vind dat daar oor tyd 'n daling van proteïen hoeveelheid was, met die uitsondering van 'n paar wat te maak het met patogenese. Wat interessant was, was dat die dormant reaksie proteïene oor tyd ook gedaal het in vergelyking met pH 7. Dit wys dat die grotendeels die reaksie tot stres 'n daling is van die sel se proteoom reaksie. Dit was ook voor die hand liggend dat die tipe VII sekresie sisteem 'n groot rol speel in die stres response. Dit was moontlik om te sien in beide die proteïen hoeveelheid en na-translasionele modifikasies. Die proteïen omset bepaling in *M. tuberculosis* was 'n groot verrassing. Daar was byna geen opname van leusien tydens die stres nie. Daarom kon ons nie bepaal wat die verskil in omset tussen die stres reaksies was nie. Terwyl dit jammer was dat ons die geleentheid mis geloop het om verskille in suur stres te bepaal met betrekking tot omset was dit interessant dat dit die uitkoms was. Dit beteken dus dat die bakterieë vinnig iets belangrik

soos aminosuur opname afsluit wat naamlik vra stel oor die rol van dié sisteem tydens infeksie.

In **hoofstuk 4** gebruik ons genoom data van kliniese *M. tuberculosis* stamme wat vanuit pasiënte afkomstig is. Ons het begin deur om te kyk hoe volop die *ppe38* mutant is in verskillende geslagslyne van *M. tuberculosis* is. In hierdie eksperimente het ons gevind dat die meerderheid van *ppe38* uitslaan mutante in geslagslyn twee voorkom en tot ‘n minder mate in geslagslyn vier. Met verdere inspeksie in die breekpunte van die *ppe38* mutant kon ons ‘n verskil in die DNS volgorde vind. So was dit duidelik dat in sekere gevalle in die geslagslyn 4 stamme die mutant kon hervorm. Ons het dit opgevolg deur om ‘n program te skryf wat hierdie tipe breekpunte in die DNS volgorde kon identifiseer. Met hierdie informasie kon ons meer breuke vind wat lei tot ‘n hervorming van gene in die genoom. Dit was verder gebruik om ‘n proteomika databasis op te stel en sodoende kon ons die proteïene vind wat ooreenstem met die hervormde gene. ‘n Noemenswaardige voorbeeld was die *Rv2323-Rv2628* hervorming in geslagslyn twee. Ons kon dus hier bewys dat in die afwesigheid van horisontale geen deling, kan *M. tuberculosis* nog steeds nuwe genetiese materiaal skep deur om bestaande genetiese materiaal te gebruik.

In **hoofstuk 5** word die ondersoek in *M. tuberculosis ppe38* verder gevat. Hierdie hoofstuk behels die ondersoek van die *ppe38* mutant en die effek op die aangebore immuun sisteem, spesifiek die makrofaag. Met die gebruik van massa spektrometrie het ons ontleed dat die makrofaag ‘n laer reaksie toon tydens infeksie van die *ppe38* uitslaan mutant. Daar word verder gewys dat dié response gebeur as gevolg van ‘n wanbalans in die Kernfactor kappa B proteïne. Met konfokale mikroskopie het ons die pad in diepte bestudeer en kon ‘n verskil in translokasie van Rel proteïene vind. Dit lewer interessante resultate wat die inflammatoriese kapasiteit van die immuun sel beïnvloed. Opvolg werk met ELISA toetse het ook bevestig dat anti-inflammatoriese faktore meer volop is in die supernatant van selle geïnfekteer met die uitslaan mutant. Dit is dus moontlik dat ‘n verlies in sekresie van die PE-PGRS proteïene lei tot ‘n laer aanvanklike immuun reaksie wat oor tyd bydrae tot ‘n groter bakteriese sel telling in geïnfekteerde individu. Dit is inderdaad ‘n kenmerk van geslagslyn 2 mikobakterieë.

In **Hoofstuk 6** word pasiënte wat tydens aktiewe tuberkulose asook na die afloop van behandeling met antibiotika gevolg. Verskillende mates word geneem, onder andere die bakteriese telling, FDG PET-CT skanderings, asook alveolêre brongiale spoel (BAL) afkomstig van die pasiënt long. Ons vind hier dat die meerderheid van pasiënte geen telbare bakteria na die afloop van behandeling het nie. Daar was ook ongelukkig pasiënte met telbare bakterieë wat slegs in die BAL voorkom maar nie in hul spoeg nie.

Hierdie pasiënte word gesien as gesond in die konteks van ‘n speeksel kweek toets, maar in die long met bewys van bakterieë en FDG PET-CT is dit ‘n ander verhaal. Ons kon ook verder met proteomika sien dat een van die pasiënte sou herbesmet word met mikobakterieë maande na die besoek aan die hospitaal. Dit is dus moontlik dat pasiënte kan presenteer as gesond maar eintlik nie is nie, al is dit in ‘n klein proporsie van gevalle is dit noodsaaklik om dit na te volg. Met behulp van proteomika is dit ook moontlik om te ontsyfer watter pasiënte eintlik toon met ‘n gevaar van herbesmetting. Met verdere ondersoek is dit dus moontlik om hierdie pasiënte uit te ken voor hul herbesmetting en in te gryp as dit nodig is met antibiotika.

**Hoofstuk 7** is die afronding van al die proteomika kennis wat in hierdie tesis gebruik word. Terwyl daar baie biologiese vra beantwoord kan word, is dit alles afhanklik van die data analise wat daar agter sit. Proteomika genereer ‘n groot aantal data, wat dit amper onmoontlik maak om met meer tradisionele maniere te analiseer. Die analise is ook afhangend van die tipe eksperimente wat gedoen word. Dit is vir die rede dat programmering nodig is om goeie insig te kry in die agterliggende data. Hierdie skep ‘n probleem vir vele ondersoekers, want programmering vir data analise en gevorderde statistiek vorm nie deel van enige tradisionele opleiding nie. Om dit te voorkom en terug te gee aan die proteomika gemeenskap het ons alles wat ons van dié veld geleer het gekonsolideer in ‘n internet gebaseerde data analise platform. Hierdie hoofstuk en sy addendum is die beskryf van hierdie platform. Vanaf die publikasie datum word die platform veel gebruik deur meerdere mense met die grootste gebruikers basis in die VSA, Duitsland, Nederland en VK.

## Afsluiting

Ten slot, die aanname van tegnologie soos proteomika en genomika asook die standaardisering van die werk is besig om ‘n nuwe era te bring in biologiese navorsing. Met die korrekte gebruik van hierdie tegnologie kan daar nuwe insigte kom in sellulêre dinamika wat nog niet voorheen moontlik was nie. Die tegnologie maak dit naamlik moontlik om grootskaalse sisteme direk te bestudeer. Met hierdie tegnologie het ons gevind dat tot eenvoudige organismes soos bakterie wel ‘n hoë gehalte van kompleksiteit besit waarin nog veel geheime verberg sit. Met betrekking tot die fisiologie en evolusionêre eienskappe van *M. tuberculosis* en sy variante het ons nog baie om te leer. In hierdie studie het ons meer insig gekry in belangrike sisteme van die mikobakterieë. Dit is veral duidelik dat die bakterie in die konteks van sy omgewing of gasheer gestudeer moet word vir die beste insig in sy gedrag. Met die kombinasie van sisteem wye karakterisering en verstand van die fundamentele prosesse wat *M. tuberculosis* laat uitstaan bo ander patogene kan ons hoop om effektief te wees in die stryd teen die patogeen.

# DANKWOORD

Dit is die grote, meer dan die einde van ’n studie maar die einde van ’n tien jaar lange reis, of sal ek sê stryd. Wat ‘n sprong was dit vanaf my meesters tot my PhD, seker het ek in die tyd gegroei as wetenskaplike maar ek sou my bemis as ek nie erken dat ek ook veel gegroei het as ‘n mens nie. Hierdie vlak van opleiding kom nie sonder hulp nie, en hulp het ek in stukke gehad. Vir die allerlaaste keer het ek die voorreg om in so ‘n dokument die mense wat my gehelp het tot by hierdie punt te bedank en te eer.

Starting a PhD study is a team effort, I would like to thank my promotors who have started me on this path and stuck with me to the end. This says something about their character if I do dare say so myself. As I have abruptly switched to English, I think it is safe to say I will start with **Samantha Sampson.** Sam I remember when you joined, or shall I say re-joined, the department (as an associate professor?) back then. I was very much in awe of the new group leader who will be joining us from prestigious places as Imperial College London and Harvard. As a young researcher I had the singular goal of somehow joining your group. This is why I will never forget the day when you entered my old office, in an old building, and spoke to me about a potential PhD project with you. Who could say no to working in the top group of the department, it would be crazy to turn you down. With your leadership and care for your students I can surely say that I made the right decision in joining your group. Thank you for your support throughout my PhD and for creating a wonderful group. Mijn promotoren komen als een duo, dus wil ik ook graag een woord van dank richten aan **Wilbert Bitter.** Wilbert, jij bent een van de slimste personen die ik ooit heb ontmoet en ik voel me vereerd dat ik zoveel van jou heb mogen leren. Al heb ik de meerderheid van mijn tijd als promovendus in Zuid-Afrika gespendeerd, voelde ik mij altijd welkom in jouw lab. Jij hebt mij zowel binnen als buiten het lab geholpen in meerdere manieren dan alleen wetenschappelijk. Ik moet zeggen dat je grootste geschenk, al was dat niet de bedoeling, de introductie was tot de persoon met wie ik de rest van mijn leven zal doorbrengen. Het was een groot plezier om onderdeel te zijn van jouw team en de vele interessante projecten de afgelopen jaren.

Jy dink ek het jou vergeet né ou Heunis. My co-supervisor, **Tiaan Heunis** kry naamlik sy eie deel. Jis dude, daar is regtig nie woorde nie. Sonder jou was daar maar baie min wat sou werk in die werk wat hier gepresenteer is. Jy is meer dan ‘n co-supervisor, jy is ‘n goeie vriend. Toe ek jou eers ontmoet, jare gelede in die Department van Mikrobiologie by Stellenbosch het ek nooit gedink dat die perd wat sit en speel met sy kar battery jare later so ‘n belangrike persoon sal word in my persoonlike en professionele lewe nie. Dankie vir jou bydrae en dankie vir al die koffies, vakansies, praatjies, intellektuele

terugvoer en mal eksperimente wat ons saam gedoen het. Wie sou dink dat ‘n paar manne uit Tygerberg nou hier sou staan. Ek sien uit om te luister na al die mal goed wat jy gaan doen in jou loopbaan. Hartlik dank vir jou bydrae en jou vriendskap, wie weet miskien doen ons nog cool dinge saam in die toekoms.

Ek will ook my vriende wat ek in die werkplek gemaak het, maar vriende geword het buite ook, hartelik bedank. Bonjour **Caroline (GG)**, c’est l’étendue de mon français. Vos leçons ont vraiment aidé même si j’ai abandonné à mi-chemin. Number three of the golden trio and co-founder of the **T**iaan **I**nstitute for **T**echnology and **S**cience. Thanks for going out of your way to be such an awesome friend, colleague and collaborator. Especially thanks for the beautiful images in this thesis. **Jomien**, mevrou vloeisitometrie in lewende lywe. Dankie vir al die kere wat ek kon bydra tot jou navorsing en dat jy daar was om te luister as ek dit nodig gehad het. Dankie vir al jou insig in die werk ook. Dit word opreg waardeur. **Alex** and **Suus,** my very first office mates when I joined the VU. Thanks for helping me navigate things at the start and making me feel very welcome in the laboratory. I enjoyed working alongside both of you and my only complaint was that it was too short.

I would also like to thank my past and present **Host-Pathogen Mycobactomics lab** mates: **Nastassja, Trisha, Caitlyn, Su-Marie, Lesedi, Pamela, Bahar, Julian, Zimvo, Hanri, Namaunga, Caroline, Marvin, Antoinette, Pumla** and **Juanelle.** Every one of you contributed to an awesome working environment and interesting discussions over the years. Thanks for all the input and good luck to those who are finishing up as well. Thanks to **Ruben, Anzaan** and **Jason** for helping me out when I knew absolutely nothing about genomics. You made this journey significantly easier. Also thanks to **Jessie,** for just being who you are. You brought an air of light heartedness when it was necessary and it was always fun having you around. We really saw eye to eye in the end. I would not be where I am today if **Ian** and **Albertus** did not see it fit to take accept a young student from main campus Stellenbosch into their lab all those years ago. Thank you for that and your mentorship for the first few years. I also had many nice chats in the hallways with **Nasiema, Craig, Marlo, Liesl, Gina, Carine, Annemie, Lizma, Rika** and **Stefanie** which would brighten up my day**.**

I would also like to thank **Rob**, **Paul, Monique** and **Gerhard** for many thoughtful discussions over the years and their help in my career thus far. All four of you have helped me immensely during my time at Stellenbosch at one time or another and for that I am very grateful.

There are also my colleagues that were 10 000 km away most of the time but always fun to hang out with. I would like to thank the awesome people of the **VU/Vumc**. Thanks **Vincent, Maikel, Kinki, Louis, Catalin, Trang, Maroeska, Markus, Melanie** and **Maurice** for showing me around and helping me in the laboratory when I needed it. Especially as I was there for so short at a time, this was a life saver. I would also like to thank **Merel, Eva, Bea, Aniek, Lisette, Vien, Dung, Gina** and **Sita.** While I was no longer in the lab, the beers and quick chats provided a welcome distraction from writing this thesis. I especially enjoyed the random beers during the week.

Veel dank aan **Roy**, **Corinne, Sander, Greg** en **Bart** voor jullie directe bijdragen tot de hoofdstukken in dit proefschrift. Ook aan **Janneke**, die jaar na jaar trainingen gaf om het mogelijk te maken dat ik in het lab mocht werken. Dankzij jou snap ik helemaal hoe de autoclaaf werkt. Heel veel dank aan **Edith** voor de geweldige input tijdens de type VII werk besprekingen. Veel van jouw tips hebben een aantal experimenten mogelijk gemaakt, andere tips heb ik meegenomen naar Zuid-Afrika en die worden zeker nog steeds gebruikt. Ten slot, bedankt aan allemaal die deel vormen van de MMM, ook die nog niet waren genoemd, namelijk **Marion, Diana, Renate, Ben, Christina, Coen, Peter, Joen, Wouter, Mathijs.** Het was echt heel gezellig met jullie bij de borrels, labuitjes, kerstdiners en nog veel meer.

I would like extend a thanks to the various people who collaborated on the different chapters of this thesis but have not been mentioned yet. Thanks to **Leanie Kleynhans, Arnap Pain, Karin Schildermans, Sven Bruijns, Inge Mertens, Rouxjeane Venter, Andre Loxton, Matthias Trost, Nelita du Plessis, Jill Winter, Fanie Malherbe** and **Bavesh Kana**. Without your various inputs and contributions I would be quite a few chapters light.

Volgende wil ek ‘n woord van dank aan my familie gee, namelik **Michelle Meyer**, **Mark Gallant** asook sy vrouw **Elizabeth** en my broer **Joshua**. Dankie aan my ouers vir die fondasie wat hulle gelê het wat my gehelp het om tot by hierdie punt te kom. Ons het baie saam deur gebring, dis ‘n wonderwerk dinge het relatief goed uitgewerk. As julle mooi onthou het ek aan julle gesê toe ek vyf of so was dat ek gaan skilder in Parys langs die Senne. Ek het hom gemis met ‘n paar 100 kilometer en sit nou langs die Amstel. Sonder die waardes wat ek by julle geleer het, het ek verseker nie hier beland nie. Spesiale woord aan my moeder wat haar eie stryd deur maak. Uitsettings vermoë is hoe ons bo kom, al gaan dinge swaar kan ‘n mens altyd deur druk is iets wat ek by jou geleer het. Ek sien uit om oor twintig jaar weer die sin saam met jou te lees. ‘n Spesiale woord van dank aan my ouma, **Phillipina Francina Gallant,** wat nie meer met ons is nie. Jy het my op hierdie pad gesit en by my gestaan deur die jare toe dinge baie swaarder vir

ons was. My tyd op ons klein boerdery met jou was sonder twyfel die beste jare van my lewe. Ontsettend baie dankie vir alles wat jy vir my gedoen het, veral toe dit nie jou plig was nie. Dit was ‘n voorreg om van jou te leer en hierdie graad is net soveel joune as wat dit myne is.

Baie dankie vir julle geduld, **Phillip** en **Rainier,** wat na Amsterdam verhuis is maar konstant moet hoor dat ek besig is met my PhD en nie kan kuier nie. Julle sal bly wees om te hoor dat dit nou uiteindelik klaar is! Ek sien uit om op te vang in die kroeg of sommer vir ‘n lekker braai hier in die tuin.

Dankie ook aan **Alida van der Mescht, Arno Visser, Luka van der Merwe, Tiaan van der Merwe,** en **Wilmi Naude** wat my ondersteun het vir ‘n groot deel van my studies. Dit was altyd lekker om by julle in te val vir ‘n koffie. Al die aande in buckleys of die shak het seker bygedra tot hoe lang die affêre nou eintlik gevat het, maar ek het geen klagtes nie. Julle is lekker mense!

*Daarsy boys, Daarsy boys laat hy val waar hy wil.* Ek sal verseker nie die boys vergeet nie. Ek wil die manne wat al jare met my saam staan op die beurt sit. **Wessel le Roux, Christo Kotze, Johan Havenga, Wian Steyn en Ryno Weyers**, ons is nou almal versprei oor die wêreld, wie sou dit nou raai van ‘n klomp kak aanjaers uit die voorstede. Dit is regtig besonder dat ek op enige een van julle kan reken as dit moet. Ek sien uit na die dag wat ons weer lekker om ‘n vuur kan sit en opvang oor ‘n biertjie en ‘n lekker braai. Dit is nou die merk van vriendskap as jy my vra, dit is ‘n eer om julle vriende te noem. **Mareli Johnson,** ek onthou toe ek ook een van hierdie vir jou geskryf het in my meesters, as ek geweet het dit die laaste keer was wat jy een sou lees sou ek iets beter skryf. Jou ondersteuning oor die jare as ‘n metgesel en ‘n vriendin word opreg waardeur. Ek mis jou nog elke dag en wens ek kon dit met jou deel. Dankie vir alles, jy weet wat dit behels. Rus sag Lils.

Een heel speciaal dankwoord aan familie van Leeuwen, **Rolinka, Jorine, Mitamarian** en **Hans.** Heel erg bedankt voor al die steun tijdens mijn promotie. Ik vind het echt bijzonder dat jullie mij zo ver van huis een thuisgevoel geven. Ik vind het ook heel leuk dat ik mag mee doen aan familie en alle leuke Nederlandse tradities met jullie. Het maakt zo’n groot verschil dat ik vaak vergeet dat ik eigenlijk een buitenlander ben. Ik zie uit naar veel meer gezelligheid samen.

Lieve **Lisanne,** wat kan ik nou zeggen tegen zo’n speciaal persoon. Er zijn veel te veel dingen om je voor te bedanken, het wordt bijna weer een proefschrift. Heel erg bedankt voor alles wat je hebt gedaan, zonder jou zou ik nooit dit punt hebben bereikt.

Jou ontmoeten was het leukste deel van mijn studie! Bedankt voor je hulp, alle leuke reizen, dat je gewoon zo'n leuke vriendin bent en alles wat je voor me doet en deed tijdens de zware dagen van deze studie. We hadden tot dusver zo'n geweldige tijd samen, vooral met een echt lange-afstand relatie van 10.000 km de afgelopen 5 jaar! Gelukkig wonen we nu lekker samen in ons mooie huis met alle planten, stekjes en mooie tuin. De afgelopen 5 jaar met jou waren echt geweldig en ik kijk uit naar de komende tachtig.

# CURRICULUM VITAE

James Gallant was born on 19 November 1990 in Cape Town, South Africa. Upon graduation from De Kuilen high school in 2008 he attended Stellenbosch University for the degree Bachelor of science in molecular biology and biotechnology. During his undergraduate study, he was recruited to work as a research assistant in the Department of Microbiology at Stellenbosch University. At this time he acquired direct experience in a research laboratory and the techniques associated with molecular microbiology. As a part of his duties at the Department of Microbiology, he also tutored various practical courses to other undergraduate students gaining valuable teaching experience. James graduated from Stellenbosch University Bachelor of Science degree program with extra credits and left the Department of Microbiology in 2012. He was selected to join the competitive Bachelor of Science, honnours in Molecular Biology program in the DST/NRF Centre of Excellence in Biomedical TB research at the University of Stellenbosch, Tygerberg Campus on a partial scholarship. This was his introduction to Tuberculosis and the importance of the topic both locally and globally. He graduated top of his class with a *cum laude* passing grade in 2013 and won several competitive grants and scholarships to continue with his studies. He enrolled for the degree Master in Science in Molecular Biology in 2014 continuing on his research topic and extended the work he did previously. James started tutoring an Honnours practical course in the DST/NRF centre of Excellence in Biomedical TB research continued to do so for 4 years. James graduated *cum laude* from his Masters program after successfully defending his thesis.

In 2015 James met Wilbert Bitter and Samantha Sampson and was approached for a prospective joint PhD program between Stellenbosch University and the Vrije Universiteit Amsterdam. Under the supervision of Wilbert Bitter, Samantha Sampson and Tiaan Heunis, James started his PhD research in 2016 after being awarded the prestigious DTTP scholarship. The focus of his research was on the proteomics and genomics of *M. tuberculosis* as it pertains to the *pe/ppe* proteins including their evolution and host-pathogen interactions. His research was conducted in laboratories across South Africa, the Netherlands and Belgium.

# SCHOLARSHIPS AND GRANTS

## Undergraduate:

- Department of Microbiology Departmental scholarship (2011).

## Post-graduate Honnours:

- National Research foundation Innovation scholarship (2013).

## Post-graduate Master of Science:

- Harry Crossley Foundation project funding (2014, 2015)
- DST/NRF centre of Excellence departmental funding (2014, 2015)
- Stellenbosch merit bursary for academic excellence (2014, 2015)
- Stella and Paul Loewenstein charitable and educational trust (2014)
- NRF-MRC Health and Allied Masters and Doctoral Scholarship (2015)

## Post-graduate Doctor of Philosophy

- NRF-VU Desmond Tutu Doctoral Training Program (2016-2019)
- Stellenbosch merit bursary for academic excellence (2016-2019)
- Harry Crossley Foundation project Funding (2018, 2019)
- HD Brede award for best publication in infectious disease (2021)

# LIST OF PUBLICATIONS

**James Gallant***, Pumla Mesatywa*, Samantha L. Sampson, Tiaan Heunis. *Multidimensional Proteomic Analysis of Mycobacterium tuberculosis during Acid Stress. Manuscript in preparation. *Authors contributed equally.*

Rob J. M. van Spanning, Qingtian Guan , Chrats Melkonian, **James Gallant** *et al. A methanotrophic Mycobacterium dominates a cave 2 microbial ecosystem. Manuscript submitted.*

Eva Habjan, Vien QT Ho, **James Gallant** *et al Screening of anti-tuberculosis Compounds using a Zebrafish Infection Model identifies an Aspartyl-tRNA Synthetase Inhibitor. Manuscript accepted*

**James Gallant,** Tiaan Heunis, Caroline Beltran, Karin Schildermans, Sven Bruijns, Inge Mertens, Wilbert Bitter, Samantha Sampson**.** *PPE38-dependent substrates of M. tuberculosis alter NF-kB signalling and inflammatory responses in macrophages. Manuscript accepted in Frontiers in Immunology.*

**James Luke Gallant**, Tiaan Heunis, Samantha Leigh Sampson, Wilbert Bitter. *ProVision: a web-based platform for rapid analysis of proteomics data processed by MaxQuant*. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btaa620

Caroline GG Beltran, Tiaan Heunis, **James Gallant**, Rouxjeane Venter, Nelita Du Plessis, Andre G Loxton, Matthias Trost, Jill Winter, Stephanus T Malherbe, Bavesh D Kana, Gerhard Walzl. *Investigating Non-sterilizing Cure in TB Patients at the End of Successful Anti-TB Therapy*. Frontiers in cellular and infection microbiology 2020. https://doi.org/10.3389/fcimb.2020.00443

**James Gallant**, Jomien Mouton, Roy Ummels, Corinne ten Hagen-Jongman, Nastassja Kriel, Arnab Pain, Robin M Warren, Wilbert Bitter, Tiaan Heunis, Samantha L Sampson. *Identification of gene fusion events in Mycobacterium tuberculosis that encode chimeric proteins*. Nucleic Acids Research, Genomics and Bioinformatics 2020. https://doi.org/10.1093/nargab/lqaa033

Jacoba Martina Mouton, Tiaan Heunis, Anzaan Dippenaar, **James Gallant**, Léanie Kleynhans, Samantha Leigh Sampson. *Comprehensive characterization of the attenuated double auxotroph Mycobacterium tuberculosis ΔleuDΔpanCD as an alternative to H37Rv*. Frontiers in microbiology 2019. https://doi.org/10.3389/fmicb.2019.01922

C Sao Emani, **JL Gallant**, IJ Wiid, B Baker. The role of low molecular weight thiols in Mycobacterium tuberculosis. Tuberculosis 2019. https://doi.org/10.1016/j.tube.2019.04.003

Nastassja L Kriel, **James Gallant**, Niël van Wyk, Paul van Helden, Samantha L Sampson, Robin M Warren, Monique J Williams. *Mycobacterial nucleoid associated proteins: an added dimension in gene regulation*. Tuberculosis 2018. https://doi.org/10.1016/j.tube.2017.12.004

**JL Gallant**, AJ Viljoen, PD Van Helden, IJF Wiid. *Glutamate Dehydrogenase Is Required by Mycobacterium bovis BCG for Resistance to Cellular Stress.* PloS one 2016. https://doi.org/10.1371/journal.pone.0147706